Genomic Tools Reveal Changing Plasmodium falciparum Populations

Thesis advisor: Professor Dyann F. Wirth Rachel Fath Daniels

Genomic tools reveal changing Plasmodium falciparum *populations*

Abstract

A new era of malaria eradication programs relies on increased knowledge of the parasite through sequencing of the *Plasmodium* genome. Programs call for re-orientation at specific epidemiological markers as regions move from control towards pre- and total elimination. However, relatively little is known about the effects of intervention strategies on the parasite population or if the epidemiological cues correspond to effects on the parasite population.

We hypothesized that genomic tools could be used to track population changes in *Plasmodium falciparum* to detect significant shifts as eradication programs apply interventions. Making use of new whole-genome sequencing data as well as GWAS and other studies, we used SNPs as biological markers for regions associated with drug resistance as well as a set of neutral SNPs to identify individual parasites.

By utilizing tools developed as proxy for full genomic sequencing of the human pathogen *Plasmodium falciparum*, we characterized and tracked parasite populations to test for changes over time and between populations. When applied to markers under selection - those associated with reduced antimalarial drug sensitivity - we were able to track migration of resistance-associated mutations in the population and identify new mutations with potential implications for resistance.

Using a population genetic analysis toolbox to study changes in neutral allele frequencies in samples from the field, we found significant population changes over time that included restricted effective population size, reduced complexity of infections, and evidence for both clonal and epidemic propagation of parasites.

Contents

Listing of figures

Listing of tables

To grands and greats and those grand ones, greatly missed.

Acknowledgments

Without support and encouragement from Professor Dyann Wirth, this work and my PhD would have been impossible. I cannot thank her enough for the opportunities that she has made available for me and for the doors that she's opened. Professors Pardis Sabeti and Dan Hartl remain much-loved mentors, always willing to listen to my questions, rants, and occasional gloats and always happy to offer advice and tough love. Sarah Volkman has been a source of inspiration and guidance and a fount of wisdom both in my life and in the lab.

I am profoundly grateful for the people who dared me (literally, in the case of Kayla Barnes) to send in that first application to start me on this path. Clarissa Valim, Carmen Mejia, and Anna Andersen have helped remind me of the important things and encouraged me to follow my dreams, once I figured out what they might be.

My collaborators astound me with their intelligence, passion, and compassion. I've been fortunate to spend many hours with Hsiao-Han Chang and feel it has been a gift. Danny Milner likewise is happy to talk science, but equally happy to talk about anything at all.

I have not yet had the opportunity to directly collaborate with Amanda Lukens or Ulf Ribacke, but their input, training, opinions and guidance in and out of the lab have made me a better researcher and a better person.

I also thank the other members of the lab, both past and present, for helpful advice and wonderful conversations.

Of course, I am nothing at all without my family. My parents, Nanette and Joseph, deserve huge thanks and most of the credit. Their determination that I really did have a plan for all of this and their confidence that I would succeed in turn gave me confidence to push through my frustrations. My sister Megan continues to inspire with her own successes and life experiences shared. Their love for me helped me through tough times.

In particular I thank my husband, Noah, who gets it, and his parents Norman and Anne

for their support, interest, and love.

Friends provided much-needed relief and equally-needed sympathy. Jen Enus, Jennifer Poth, Kristin Piltzecker, and Jessica Hook remain long-time friends who haven't been scared off by my scary stressed-out phone calls, messages, and emails. Jeff and Liz Baker, Jeff Wasilko, and Evan Wieder are forever sympathetic and encouraging and refuse to accept excuses, no matter how often I try to use them.

While it may take a village to raise a child, it takes many villages both abroad and at home to try to save them from malaria. I am grateful to the collaborators and collection teams in the field sites around the world whose work contributed to mine. In particular, Professor Daouda Ndiaye and his team in Senegal have been instrumental, trusting me with their precious samples and helping me to understand the complexities and the beauties of field work in Dakar and Thiès. Sungano Mharakurwa, Dan Bridges, and Moonga Hawela have introduced me to their fantastic work in Zambia and also offered me unparalleled opportunities to learn about malaria in their country.

Much of the content of this body of work has been published in peer-reviewed media and used with permission. Chapter 2 appears as 'Population genetics and parasite diversity' in *Evolution of Virulence in Eukaryotic Microbes* (D. Sibley, ed.) from Wiley-Blackwell 2012. Hsiao-Han Chang and I co-wrote this chapter with the guidance of Dan Hartl.

Chapter 3 is adapted from 'Rapid, field-deployable method for genotyping and discovery of drug-resistance single nucleotide polymorphisms in *Plasmodium falciparum*' in *Antimicrobial Agents and Chemotherapy*, 2012 Jun;56(6):2976-86. Epub 2012 Mar 19. I planned and performed all experiments and analyses and wrote the manuscript under the guidance and with the advice of Daouda Ndiaye, Papa Diogoye Sène, Pardis Sabeti, Sarah Volkman, Soulemayne Mboup, and Dyann Wirth. Mikeal Wall and Jason McKinney assisted with much-needed advice and help with development and optimization of the assays.

The material of Chapter 4 appears in part in 'Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal' published in *PLoS ONE* 8(4): e60780. doi:10.1371/journal.pone.0060780. Papa Diogoye Sène and I extracted samples collected from the site under the supervision of Daouda Ndiaye and Soulemayne Mboup. Hsiao-Han Change, Danny Park, and I performed analyses with insightful comments from Stephen Schaffner, Amanda Lukens, and Daria Van Tyne. The manuscript was written by Hsiao-Han Chang and myself with assistance from Dan Neafsey and under guidance of Dyann Wirth, Pardis Sabeti, Dan Hartl, and Sarah Volkman.

Chapter 5 is based on 'Human cerebral malaria and *Plasmodium falciparum* genotypes

in Malawi', published in *Malaria Journal* (2012, 11:35). Working with autopsy samples extracted by Jimmy Vareta, Jacqui Montgomery, and Danny Milner, I genotyped all samples and performed initial analysis with assistance from Clarissa Valim. Clarissa performed additional statistical analysis and the manuscript was written and other tests were performed along with fantastic commentary and suggestions by the other co-authors of the paper.

1

Introduction

At present, approximately half of the world's population lives at risk of malaria infection. The World Health Organization (WHO) reports that more than 200 million infections and 600,000 deaths from malaria in 2012. The deaths are attributed primarily to pregnant women and to children under the age of 5 whose immune systems are still naïve to the parasite [125].

Beyond the devastating death rate due to *P. falciparum* malaria, the continued economic impact of infections is enormous, with government costs for health clinics, supplies, and workers as well as lost tourism revenue. In addition, individual costs of missed economic opportunities, time out of work due to illness, and the financial burden of drugs and treatments all add up to decreased productivity in already poverty-stricken regions of the globe [34]. Treatment and control of malaria remains a priority of both governments and international organizations.

While historically found in tropical and sub-tropical regions of the globe, the blood-borne parasite currently exists in greatest prevalence in Sub-Saharan Africa, where more than 80%

of malaria cases occur. *Plasmodium falciparum* is typically known above the other members of the human-infecting *Plasmodium* family for its global mortality impact [125].

1.1 Parasite biology

Plasmodium is part of the phylum Apicomplexa, whose members are characterized primarily by an apicoplast organelle integral to the intracellular parasite's strategy to infect its vertebrate host's cells. This eukaryote employs a complex life cycle requiring not only vertebrate hosts, but also insect vectors. Upon biting its host, the mosquito or fly vector releases *Plasmodium* sporozoites from its salivary glands. This motile haploid form traverses from the peripheral blood vessels to the host liver. Here the sporozoites differentiate within hepatocyte cells and multiply exponentially to emerge as merozoites to infect erythrocytes in the blood stream. Development continues as the parasite co-opts the blood cell to feed the mitotic division of a new generation of sporozoites that differentiate into trophozoites and merozoites that burst from the erythrocyte for a new round of red blood cell invasion.

During this cyclical invasion of, development within, and lysis of the host red blood cells, factors are released that signal a subset of this population to instead differentiate into male and female gametocytes that also circulate in the blood stream. Rather than infecting red blood cells, however, these gametocytes are taken up by an invertebrate host. This next stage of the life cycle is the only point at which meiosis occurs, when male and female gametocytes recombine in the insect midgut before again becoming haploid and sequestering in the salivary glands for release into the next vertebrate host when the insect takes a blood meal (Figure 1.1.1[105]).

Based on the plentiful members of the family infecting a variety of vertebrates, including avian, reptilian, and mammalian hosts as well as the relative diversity of invertebrate host species, *Plasmodium*'s origins are ancient. Its history within humans, however, is relatively recent, though still controversial in it details. Compelling evidence exists for origins ranging from 5 million to 10,000 years ago, with zoonotic transfer claimed from chimpanzees, bonobos, gorillas, and even the flat-nosed monkey [52, 58, 63, 96, 99]. While debate still exists about the details of the zoonotic leap from higher primates to humans, what is clear is that the parasite has co-evolved with its human host for a very long time. Several red

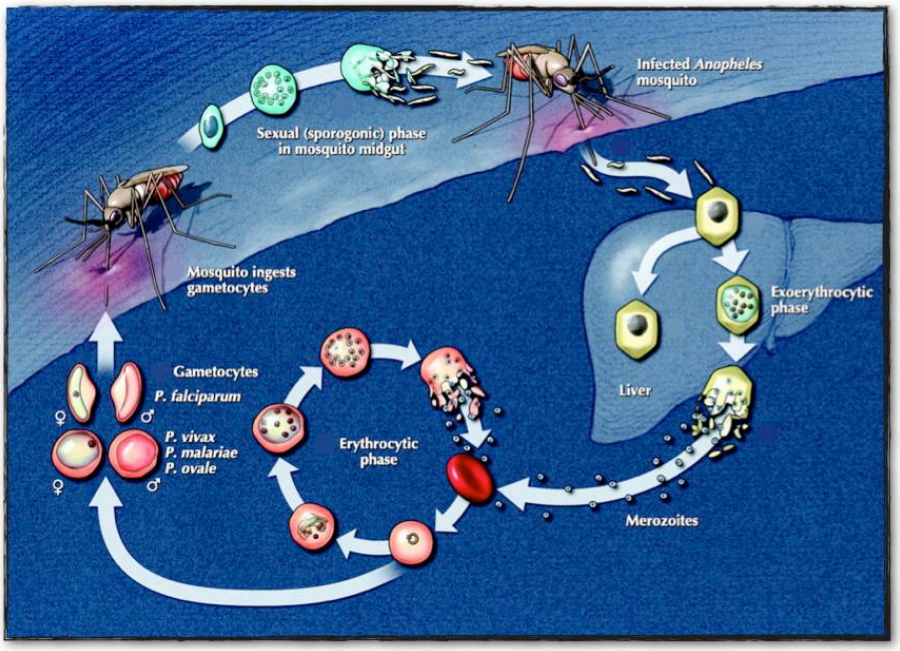


Figure 1.1.1: Life cycle of *Plasmodium* spp.

bood cell mutations have been correlated to resistance with *Plasmodium* growth or invasion. A mutation in the human hemoglobin gene results in a misformed actin complex within the red blood cell, resulting in the characteristic sickle shape associated with sickle-cell disease. The parasite is unable to successfully establish an infection in these cells. Similarly, a mutation in glucose-6-phosphate dehydrogenase (G6PD) blocks parasite development by establishment of peroxide-induced hemolysis that also stymies parasite development. Both of these mutations are most prevalent in geographical regions from which malaria originated and continues to persist in greatest concentration.

Several *Plasmodium* species preferentially infect humans: *P. ovalae, P. malariae, P. vivax*, and *P. falciparum*. Of these, the last two are largely responsible for the morbidity and mortality from malaria, the anemic disease caused by infection by these parasites. *Plasmodium vivax* is found primarily in Asiatic regions around the globe. It is less responsible for acute and fatal infections compared to its more virulent cousin *P. falciparum*, instead developing a unique stage within the host liver, the hypnozoite, that can differentiate and develop into malaria infection months and even years after the initial infection.

1.2 History and control efforts

Western medicine first described *Plasmodium* parasites as the causative malaria agent in 1880; however, ancient Chinese medical texts from more than 4,000 years ago also detail symptoms and treatment of the disease [16]. Efforts to curb the spread of the disease also have a long history, culminating previously with the Global Malaria Eradication Programme (GMEP) first instituted in 1955, with an ambitious goal of total eradication of malaria based on successful localized efforts in North and South America and armed with the newest chemical breakthrough of DDT (dichlorodiphenyltrichloroethane). While locally successful in a number of regions, the Program encountered many challenges to its success and eventually backed away from its goal of eradication to a more passive localized malaria control program by its end in 1969 [83].

Many confounding factors, from systemic lack of combined strong central organization or empowered local programs, combined with the faculty of both the vector and parasite to develop resistance to chemical interventions, led to the eventual decline of the GMEP as regions previously declared success cases for elimination suffered re-emergence and increasing transmission and enthusiasm for the Programme waned.

Despite the unraveling of the Programme's goals for total global eradication, smaller efforts continued to control and treat in localized areas.The Programme's failure led to increased awareness of the need for more research into the molecular mechanisms of the parasite and its life cycle to better understand how *Plasmodium* had escaped pressures. Finally, in 2007, the WHO and other organizations again took up the eradication banner, this time armed with more intimate knowledge of the inner workings of the parasite and with the hard-earned lessons of the previous campaign.

A coalition of several organizations, including the Malaria Eradication Research Agenda (MalERA) initiative and the WHO along with the Roll Back Malaria (RBM) Partnership have published new models for global malaria elimination. These models move away from the previous single universally applied vertical approach to instead encourage flexible intervention and surveillance strategies.

In these initiatives, each country is categorized according to its relative impact on the global malaria burden. Countries may be assigned to control intervention levels if their high

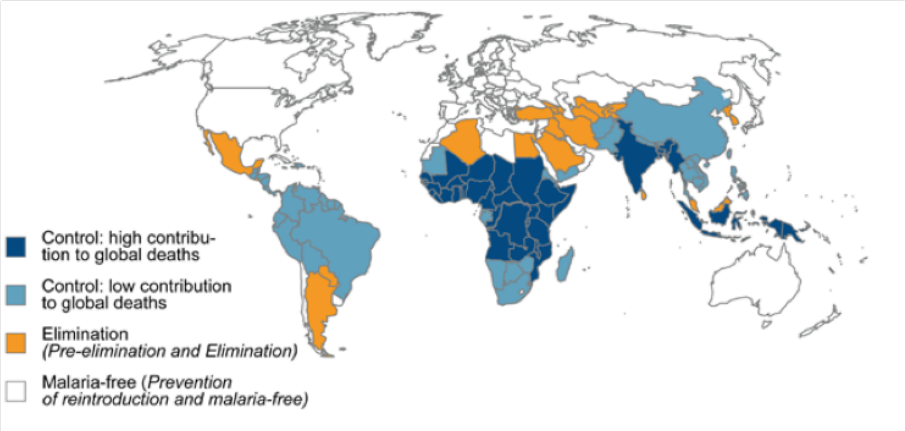


Figure 1.2.1: World Health Organization country categorization. This measure is based on individual country control status and contribution to global malaria mortality

transmission levels or numbers of infections contribute to the annual mortality due to malaria deaths. When this impact has been reduced, countries enter pre-elimination and elimination status with a goal towards country-wide elimination and a shift to surveillance measures. These categorizations allow for mobile and changing strategies to address evolving concerns in treatment, tracking, and surveillance of the parasite and its *Anopheles* mosquito vectors (Figure 1.2.1).

The WHO has further refined the definitions for each of these stages in order for individual countries to apply for an official WHO certification of malaria elimination. Individual countries are encouraged to participate on every level, from individual education to healthcare and government health ministries to plan strategies with the assistance of international collaborative teams representing a variety of clinical and epidemiological disciplines. Under this scheme, program reorientation is specifically called for as a country moves into pre-elimination status and again when no local cases are detected. Again, these specific epidemiological milestones are necessary to best make use of limited government resources to best address changing priorities for each stage. (Figure 1.2.2).

RBM's Global Malaria Action Plan (GMAP) has established milestones for global malaria control based on current intervention strategies in a hope to reduce the number of deaths from malaria. By 2010, the goal had been to reduce the number of malaria-attributed deaths by 50% from the numbers reported in 2000 to 500,000. Also by 2010, nearly every pregnant

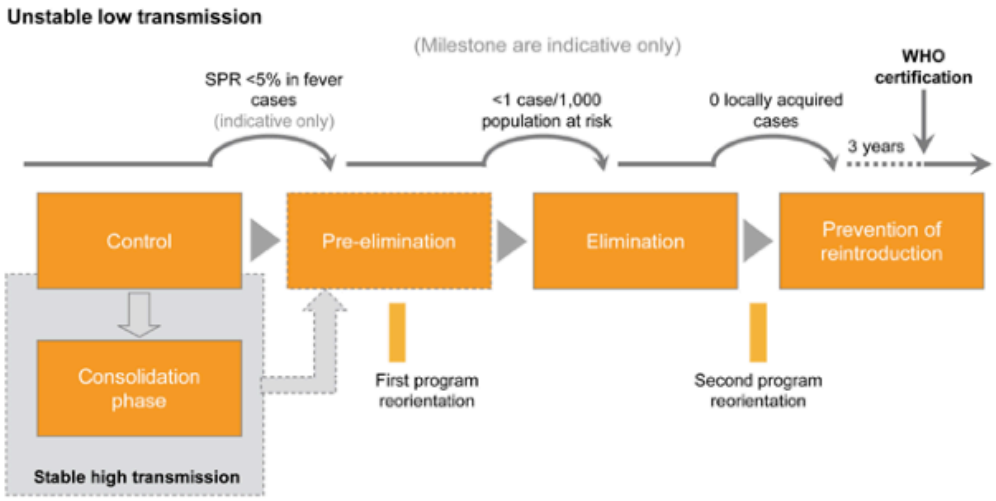


Figure 1.2.2: World Health Organization eradication milestone. World Health Organization markers towards certification as malaria-free

woman in high-transmission areas would have access to intermittent preventative treatment.

More significantly, by 2015, in addition to extending the preventative and treatment coverage to be nearly universal with accompanying reduction of malaria deaths, 8-10 countries are expected to report zero local transmission of the malaria parasite [67]. These goals would be accomplished by ensuring that 80% of people living in endemic areas have access to appropriate vector control strategies like insecticide-treated bed nets and indoor residual spraying as well as environmental control methods such as reduction of standing water and application of insecticides to mosquito breeding grounds. Similarly, this population would have access to diagnostic tests and treatments. Unlike the previous Programme efforts, the initial consolidation stages would gradually ease into control efforts. The initial stages, termed Scale Up for Impact (SUFI) require massive initial outlay and labor for materials and distribution to ensure the greatest coverage possible at the outset [68].

Also in contrast to the GMEP, this new elimination campaign requires integration of both laboratory and field clinic researchers [69]. Both MalERA and RBM call for funding of research and outline its utility throughout all stages of malaria control and eradication

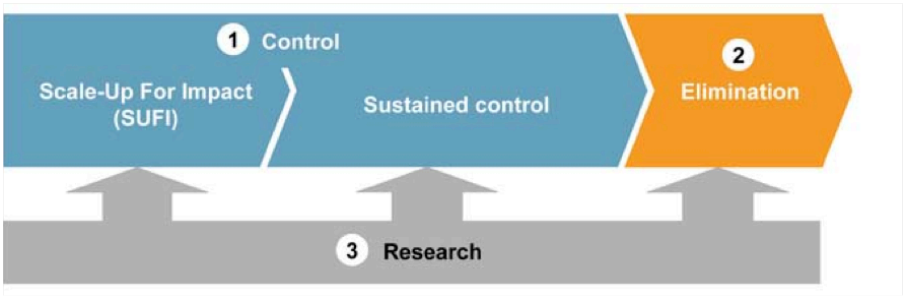


Figure 1.2.3: Global Malaria Action Plan elimination strategy. GMAP's global action plan, calling for continued research accompanying the control and elimination phases

(Figure 1.2.3).

Research efforts cover all aspects of the parasite life cycle and development to increase knowledge of the molecular machinations of the parasite, especially those involved in escape from host immune systems and development of drug resistance. In addition, sequencing a number of individuals from a variety of populations offers additional data to address the challenges inherent to the proposed elimination strategies. Using this information, fundamental and necessary tools can be developed to study the spatial and temporal distribution of both parasites and their vectors, how they are transmitted, and how changes in these characteristics effect clinical burden due to malaria infections. Further, they will help to determine if the epidemiological milestones correlate to parasite population changes indicative of reduced genetic diversity and complexity of infection through reduced meiotic outcrossing.

These new tools offer different views of changing populations and population structure in response to natural and artificial selection pressures, especially as regions move from control towards elimination. In these stages, parasite population epidemiology, diversity, and distribution are expected to change dramatically. Current gold-standard technologies like rapid diagnostic tests (RDTs) cannot detect low levels of parasites or assess the changing dynamics of effective population size or genetic diversity. They offer limited information in the face of increasing case numbers to differentiate between newly emerging infections from undetected reservoir populations, or drug treatment failures from emerging infections, or importation of parasites into a previously-cleared region. These are questions imperative to assess the effectiveness and status of control programs. This work presents several new tools developed to not only answer these questions, but to offer real-time utility to quickly inform these efforts.

1.3 Genomics

The sequence of the *Plasmodium falciparum* reference genome was first assembled in 2002 [36], catalyzing research efforts to understand parasite biology. Since publication of the original genome, isolates from a variety of geographic origins have been sequenced, revealing a diverse and motile genome [51, 70, 79, 90, 113], with antigenic variation, microsatellites, gene copy-number variations (CNVs), and single-nucleotide polymorphisms (SNPs) among the strategies that the parasite employs to rapidly adapt to natural and artificial selection.

Researchers can use these genomic markers to characterize *Plasmodium* populations in several ways. First, by using markers under selection, such as genes for parasite surface antigens, studies can determine relationships between immunity and transmission. These relationships are especially interesting as malaria vaccines are tested and clinical trials completed. In addition, SNPs and microsatellites related to reduced drug sensitivity can reveal evolutionary history of selection events, detect emergence of new mutations, and track the spread of these SNPs across the globe. Surveillance of these SNPs allows control programs to better tailor their anti-malarial drug regimens and to respond to new resistance threats before they are introduced or emerge in populations. Other markers that appear to be under positive selection are also candidates for regions under selective pressure.

Using selectively neutral markers in the genome, each of these measures of genetic diversity can also describe parasite population structure [6, 18, 86, 98]. With sufficient data from a variety of populations, researchers can measure genetic diversity between and among populations. Polymorphic markers also define linkage disequilibrium (LD), the nonrandom association of alleles. Further, these variable markers make useful tools to consider complexity of infection (COI), a metric of the number of genomes present in a patient sample. Combined, these traits characterize individual populations and can inform genetic relatedness between populations. The combination of population structure and relatedness of parasites within geographical regions allows identification of parasite types and their likely movement within and beyond these regions. This knowledge leads to better predictions of the likelihood of successful elimination [76], and thus better prioritization of limited resources.

Understanding changes in population structure of considerable importance as countries and regions move from control towards pre-elimination status. Under these conditions, the number of reported cases is considerably reduced and new surveillance strategies must be im-

plemented. The gold-standard for malaria detection, RDTs, require a minimum parasitemia in patients; when treatment programs have sufficiently reduced the burden of malaria, the number of false-positive instances is expected to increase. PCR-based methods of detection are an option for increased sensitivity, but thus far, strategies for implementation as part of control and elimination efforts have not been developed to adequately assess the true prevalence of the parasite in these low infection settings.

However, in order to be prepared for these program reorientations, it is essential to track how close populations are getting to these levels. While many studies describe extant populations and compare their population structure and other characteristics to others, few have tracked changes in the population genetics of the parasite. It is of paramount importance to detect and characterize changes in the parasite populations as regions move along the spectrum from control towards elimination, particularly as the parasite population responds to control and eradication efforts. In regions of high parasite transmission, it is common to find evidence for multiple parasite genomes within a single patient (high COI). In contrast, in areas with very low reported transmission such as areas well into pre-elimination or experiencing imported or other small-scale epidemics, either a single or limited number of *P. falciparum* genomes are present (low COI). What is unclear and so far unreported is the route between these population states and what effects the number of genomes within patients (single versus mixed infections and relative COI) has on population structure.

As transmission – and presumably the number of reported malaria cases – decreases, the overall genetic diversity will also decrease as the opportunity for meiotic crossover events between genetically distinct parasites within the mosquito midgut will become more rare. This will be co-incident with a reduced complexity of infection. With fewer parasite types in the population, crossover events that occur will be between identical or nearly-identical individuals, increasing the inbreeding levels within the population. However, several groups have reported continued transmission of mixed infections even in low transmission areas, hinting that the mechanisms for parasite transmission may be complicated by persistence of mixed gamete transmission within mosquitoes [13, 28]. This continued elevated COI could confound efforts to determine eradication status.

Identification of markers of the status of these populations will develop a better understanding of the changing population genetics and ultimately lead to better strategies and more effective distribution of limited resources in the global effort to eradicate malaria.

These measures and their implications for malaria control efforts are discussed in greater detail in the following chapters.

1.4 Summary of chapters and aims

The work of this dissertation is to utilize genomic information to better understand both natural changes in *Plasmodium falciparum* population genetics as well as those in geographical regions undergoing intense control efforts. The most informative methods to generate complete data for comparison of populations and identification of markers remains deep (high coverage) sequencing and large-scale Affymetrix SNP arrays. Using these methods, a variety of population characteristics can be described as above and in the work by Chang and Park to show population demography as well as additional sites of selection in the genome [3, 17, 91]. While sequencing entire parasite genomes continues to become more tractable in cost and computation, it remains prohibitive and excessive for large-scale population surveillance and cumbersome for results to be generated in a time frame to allow real-time assessment of the effects of control efforts on parasite populations. The feasibility of large-scale studies utilizing whole-genome sequencing is questionable, especially when other methodologies could be used to answer questions relevant to population surveillance such as changes in transmission type and frequency, complexity of infection, and allele frequency and its related variance effective population size.

We hypothesized that a limited number of genetic markers could be used to track temporal changes in parasite population genetics, especially in response to enhanced malaria control efforts. We review these principles of population genetics as they apply to *Plasmodium falciparum* parasite diversity in Chapter 2.

Toward this end, we have developed tools designed for easy implementation in a variety of research settings, then set about applying these tools to track parasite populations to test our hypothesis that significant population-level changes were happening and could be tracked in the parasite population. Building on existing knowledge of SNPs associated with reduced drug sensitivity, we designed and implemented a set of assays meant for easy genotyping in a variety of settings. In Chapter 3, we develop and apply new tools with increased sensitivity and utility to discover temporal trends in markers under selection, identify emerging or imported drug resistance mutations, and track imported resistance reported elsewhere as it

moves through the population.

An additional tool that we have developed is the *molecular barcode*, a set of SNP markers from across the *Plasmodium falciparum* genome. Using the genotypes from 24 individual alleles from the parasite's 14 chromosomes, we are able to uniquely identify individual parasites and also mixed infections [24]. We test the hypothesis of changing population structure in Chapter 4 by utilizing the neutral markers of the molecular barcode as a proxy for entire genome sequencing to track radical changes in a population over 6 years to find evidence of clonal and epidemic expansion of certain parasite types, extremely small variance effective population size, and persistence of parasite types across dry seasons.

Chapter 5 uses the molecular barcode to track parasite populations within children suffering from cerebral malaria, a severe condition that commonly results in death. Here we present a trend that single parasite types over-run the patients, rather than an accumulation of multiple parasites that overwhelm the immune system.

These studies track parasite population genetics through both neutral markers and those under selection. Using these markers in geographical populations and even within individuals can inform eradication programs of the effectiveness of their interventions.

2

Population genetics and parasite diversity

In population genetics, evolution occurs when the frequency of alleles in a population changes over time. A single organism cannot undergo evolutionary change, but individuals possess genomic diversity that collectively forms a pool of genetic variation upon which the forces that lead to evolution can act. Some authors prefer the term evolution to refer only to changes in allele frequency that result from natural selection; however, in most cases the underlying cause or causes of allele-frequency change cannot easily be assigned with certainty; however, we can develop frameworks to approximate the observed frequency changes. In this chapter, we describe an idealized model of allele-frequency change and discuss the ways in which evolutionary forces can cause these changes to differ from the model. Differences between actual data and model predictions open a variety of approaches to drawing inferences about the evolution of microbes through analysis of genetic data.

2.1 Mutation

Mutation encompasses a wide range of changes from substitution of single nucleotides to relocation of entire segments of chromosomes. Single-nucleotide polymorphisms (SNPs) include transitions and transversions that exchange bases within or across purine and pyrimidine classes. Some nucleotide substitutions can affect the codon for an amino acid without changing the amino acid (a synonymous SNP). When the mutant codon does code for a different amino acid, the SNP is nonsynonymous. Some mutations insert or delete one or a small number of nucleotide bases and shift the reading frame of the translational machinery, causing a usually non-functional product to be translated from the RNA transcript.

On a larger genetic scale than single-nucleotide changes, copy number variations (CNVs) are mutations caused by mobile genetic elements like insertion sequences, transposons, or retroelements. Genes can also be duplicated through chromosomal mutations such as transpositions and translocations; like the other types of CNVs, these mutations place genes out of their normal genetic context and away from their normal regulatory elements. The new copies may also be less prone to correction by processes such as gene conversion, leaving them free to accumulate further mutations. A special class of copy number polymorphisms consists of minisatellites and microsatellites with variable numbers of tandem repeats of relatively short stretches of DNA.

Mutation can have significant consequences on populations. For example, the effects of nonsynonymous SNPs and CNVs have been associated with reduced drug sensitivity in *Plasmodium falciparum*, the apicomplexan parasite that causes most malaria deaths. The implications of these and other types of mutations in populations of *P. falciparum* will be discussed in greater detail in later chapters.

At the beginning of population genetics, the only way to observe mutations was through direct polymorphic evidence: Mendel's vaunted peas, Kettlewell's melanic moths, and Galton's amazing catalog of continuous traits all owed their phenotypic variation to different versions of genes where mutations had caused the organisms to appear different from each other. Increasingly sophisticated techniques such as allozyme (also known as allelic isozyme) and restriction fragment length polymorphism (RFLP) analyses allowed more direct studies of protein and DNA differences between individuals within populations and among species. These early experiments formed the basis for many current theories in population genetics

and remain informative to the present day. The recent emergence of cost-effective technologies such as DNA and protein arrays has changed the scale and accelerated the discovery of diversity from single genes or proteins to genomes and proteomes in massively multiplexed studies involving larger numbers of individuals within and among populations. Advances in sequencing technologies have further enhanced the deluge of new data available to researchers. With lowered costs and increased coverage, entire genomes can be sequenced for a complete picture of polymorphic sites across populations and time. Further technological advances like hybrid selection and single-molecule amplification mean that pathogens and other microorganisms including eukaryotic microbes that cannot be separated from host material or cultured in sufficient quantities to allow easy sequencing under previous technologies can now be sequenced directly from patients or the environment. These technologies do not rely on in vitro culturing systems and offer researchers glimpses of the population genetics of pathogens whose mechanisms of virulence may be intractable for study in model organisms.

2.2 Genetic diversity and random genetic drift

Knowing the extent of genetic diversity allows us to make comparisons among populations in space or time. Studies of genetic differences among populations and the mechanisms that bring them about can reveal evolutionary history and can predict future evolutionary events. Genetic diversity is measured in several standard ways, so comparisons can be made within and among populations. Heterozygosity, originally developed to quantify allozyme differences, calculates the proportion of loci that are heterozygous. This flexible and straightforward metric allows testing between variable locations in a single diploid organism, among individuals in a population, and between species. With access to genetic sequences, other measures of genetic diversity are more commonly used. One measure, nucleotide diversity (typically symbolized by π), is the average proportion of nucleotides that differ between any randomly sampled pair of sequences. A second measure of DNA sequence variation is the number of segregating sites (usually symbolized by S)[54]. While cataloging the magnitude of genetic differences offers ways to compare populations, it does little to inform evolutionary history or relationships among populations. In order to recognize and understand the forces that change the frequency of alleles in a population, it is convenient to compare an actual population with an idealized model population in which only well-defined evolu-

tionary forces are at work. In the simplest models, the ideal population is infinitely large, mating is random, and generations do not overlap. In the absence of mutation, migration, and selection, the allele frequencies remain constant generation after generation, and with only two alleles at a locus, the genotype frequencies in a diploid organism are given by the familiar Hardy-Weinberg principle as $p^2, 2pq,$ and q^2, where p and q are allele frequencies and $p + q = 1$.

The next level of complexity still in an idealized model is to allow mutation to take place and to assume that the population is finite in size. The finite size allows random fluctuations in allele frequency to take place from generation to generation. The mutations are assumed to have negligible effects on survival and reproduction, which is called selective neutrality. Under these assumptions, the idealized model is known as the neutral model and often serves as a null hypothesis for comparison with observed data. Tajima's D is one test statistic that we can use to test differences between the observed and idealized situations. Tajima's D [106] can be easily calculated from sequences of selected loci of many individuals in a population through measures of sequence variation π and S/a, where

$$a = \sum_{n-1}^{i=1} \frac{1}{i}$$

and n is sample size. Rejection of the neutral model, where the observed value is significantly different from the expected value of Tajima's D (which is 0), can identify genes in which forces such as natural selection may be important; however, departures from neutrality can also result from the departures from any other assumptions made in the idealized model such as population growth.

In a finite population, random genetic drift changes allele frequencies over time through stochastic sampling variation that can change the relative representation of an allele in the next generation. Gametes from each parent that successfully combine to form offspring are a subset of the total possible from each parent; which particular set of alleles combines with those from another gamete is entirely random. Also random is the chance that an individual carrying a particular allele will survive and successfully reproduce. A formal model for the effects of random genetic drift was proposed independently by Wright and Fisher [1, 126] and hence is known as the Wright-Fisher model. The key simplifying assumptions are a finite and constant population size (N), nonoverlapping generations, and equal likelihood of reproduction of each individual. In its simplest form, the model also assumes no new muta-

tion and deals only with those that already exist in the population. With these assumptions, the probability that the copy number of an allele will be k in a diploid population in the next generation is

$$\frac{(2N)!}{k!(2N-k)!}\, p^k q^{2N-k}$$

where p is the frequency of the allele in the current population and $q = 1-p$. The formula here could be further used to predict the influence of genetic drift and population dynamics, such as the expected frequency of mutations in a population or the expected time for an allele to become fixed or lost by genetic drift.

Once a mutation enters the population, its possible fates are limited to either survival or extinction. With no selection, its fate is determined by purely random fluctuations in allele frequency (random genetic drift). A new mutation in a diploid population of size N has an initial allele frequency of $1/(2N)$. The odds are greatly against the allele establishing itself in the population. For a large population, the probability that the allele will be lost in the next generation is 0.368. The odds against surviving to the second or fifth generation after introduction are 0.532 and 0.732, respectively. On average, an allele that is destined to be lost will do so in approximately $2ln(2N)$ generations. Although the odds are strongly against any new neutral mutation remaining in the population for very long, there are so many new mutations that some of them do become established, and a few even drift to fixation. In the neutral model, the probability that an allele eventually becomes fixed is equal to its frequency in the population. Fixation of a new neutral allele takes a long time, however. A mutation at initial frequency $1/(2N)$ that is destined to be fixed requires an average of $4N$ generations for this to occur [55].

Through the accumulation of mutant alleles and their random changes in a frequency due to genetic drift, an equilibrium is eventually established with predictable values of π and S and other population parameters, as well as a predictable probability distribution of allele frequencies (called the allele frequency spectrum). It is against these predictions that actual data are compared to reveal genes that deviate from the assumptions of the model. For example, deviations caused by selection for biological attributes or drug resistance would be of interest.

2.3 The neutral theory

Kimura formalized the Wright-Fisher model in the neutral theory, which proposes that the primary cause of evolution in populations is through the mechanism of genetic drift on neutral alleles [53, 56]. Kimura developed this theory based on thinking at the time that the human genome might contain at least a million genes. With the degree of polymorphism of protein-coding genes then known, most segregating mutations must be selectively neutral and must confer no apparent advantage to individuals with different alleles; otherwise, the genetic load from segregating mutations would be so large that the population could not survive. Kimura also posited that even if non-neutral mutations emerged, the vast majority would be deleterious and quickly lost from the population through negative selection. One strong example used to defend this theory was the degeneracy of the codons mapped to amino acids. Kimura hypothesized that most gene products that are found in the possible genetic landscape are at some form of an optimal state and that most changes are simply variations on a theme that has been successful for many generations. In this theory, the rate of loss of neutral alleles is balanced by the mutation rate (μ) of these same neutral alleles and variation is maintained in the population. We now know that in eukaryotes with large genomes, the vast majority of the DNA does not code for proteins, and hence, the neutral theory has been extended to apply to most noncoding sequences. Because of its simplicity and ability to make specific predictions about such things as the allele frequency spectrum, the neutral theory has served as a critical null hypothesis in molecular population genetics and genomics.

2.4 Mutation and selection

The obvious alternative to the neutral theory is natural selection. First proposed by Darwin and subsequently modified to reflect updated scientific knowledge, this theory acts on individuals to change the characteristics of an entire population. Under natural selection, in contrast to the neutral theory, selection on allelic variation nonrandomly changes allele frequency owing to differences in the relative fitness of the individuals. Fitness is measured by survivorship and fecundity the ability to survive to reproduce, the number of successful matings, and the number of off-spring produced per mating. Positive selection increases the likelihood of an allele passing to the next generation through increased fitness, while negative

selection is deleterious to these chances. There are three general types of selection that work to change allele frequencies in distinctive ways. First, directional selection favors one extreme variant over the other genotypes and pushes the allele frequency toward one homozygote. If the new mutation is favored, the directional selection is positive; if the original type is favored over the mutation, negative or purifying selection occurs. Next, balancing selection chooses an intermediate solution over extremes. A well-known example of balancing selection is the fitness advantage of having one copy of the sickle-cell mutation to help protect against severe illness caused by *P. falciparum* malaria. Homozygous nonmutant genotypes are more likely to suffer severe infections, but homozygous mutant genotypes have severe anemia. The heterozygous genotype, therefore, has an advantage over both homozygous genotypes. The third general type of selection is diversifying selection, which emphasizes maximum allelic diversity. Many genes in the immune system and pathogen virulence factors that interact with the immune system are thought to be under diversifying selection.

Even when a beneficial mutation appears in a population, random genetic drift plays an important role in determining its fate, especially in the early generations. For a new mutation with a selective advantage of 1%, for example, the likelihood that the mutation will be lost in the first, second, or fifth generation is only slightly smaller than under a neutral model: The theoretical probabilities of loss when N is large are 0.364, 0.526, and 0.725, respectively. When the selective advantage s is small, the likelihood of ultimate fixation in the population is only twice the selective advantage. Selection can indirectly affect neighboring genes. When strong positive selection changes the frequency of an allele, the advantageous form sweeps through the population. If this selective sweep happens in a short period of time, neighboring alleles are carried along as genetic hitchhikers before genetic recombination can break up the allelic associations. One measure to test for the presence of selective sweeps is linkage disequilibrium, which is the nonrandom association of alleles in a chromosome. Under the neutral model, genes are in equilibrium when no particular combination of alleles (a haplotype) is more prevalent than would be expected by chance. Under selective pressure, including selective sweeps, linkage disequilibrium occurs when the frequency of certain haplotypes rises above random expectation.

Proponents of Darwin's theory of natural selection at first saw the neutral theory as being in direct opposition to the tenets of selection, and for about a decade, disagreements between the two camps appeared as polemics supporting either selection or neutrality as a universal mechanism of evolution. These grand theoretical battles are mostly in the past, and modern

researchers are more concerned with assessing the relative importance of selection and neutrality in various actual situations. In this instance, progress in sequencing technology and other aspects of genomics were important to help the field of evolutionary genetics escape the impasse. We should note here an important modification of the neutral theory known as the nearly neutral theory, which recognizes that many mutations may indeed have selective effects of their own but stresses that these effects are often small in relation to the population size [89].

2.5 Effective population size

When a population geneticist refers to population size, it is usually the effective population size that is intended. The effective size of a population, sometimes informally called the "breeding size," corresponds to the size of the ideal population discussed earlier that would have the same magnitude of random genetic drift as an actual population [126].The effective size of a population is usually smaller than the actual size and is sometimes much smaller. Symbolized by N_e, the effective population size is an important characteristic of a population that determines the effects of genetic drift interacting with selection and other processes. More formally, the effective size of a population is defined as the size of an ideal Wright-Fisher population that has the same properties with respect to genetic drift as the real population does. The effective size is often more relevant to evolutionary processes such as adaptation, including the evolution of drug resistance, than the actual population size.

The detailed consequences of random genetic drift can be measured in several distinct ways, and depending on the way random genetic drift is measured, three kinds of effective population sizes have been defined: variance effective size, eigenvalue effective size, and inbreeding effective size. Variance effective size is the size of an ideal population that would produce the same variance of allele frequency from one generation to the next as the actual population; eigenvalue effective size considers the rate of loss of heterozygosity across generations; and inbreeding effective size focuses on the probability that alleles are identical by descent.

One complication is that the three types of effective size can differ in certain demographic models. Sjödin et al.(2005) introduced the coalescent effective size and demonstrated that, if it exists, the coalescent effective size is the most general because it allows all aspects of

random genetic drift to be described by Kingman's coalescent process [102].The coalescent effective size is defined as the size of an ideal population that has the same properties of the coalescence process as an actual population. Coalescence is the convergence of the ancestral lineages of two alleles to a common ancestral allele. In an ideal population that follows the Kingman coalescence process, the time for two lineages to coalesce follows an exponential distribution with mean $2N$, and hence in an actual population, the coalescent effective size can be estimated as half the reciprocal of the rate of coalescence. Coalescent effective size does not always exist: It exists only when the factors that affect N_e take place on different timescales from coalescence events.

2.5.1 Factors affecting effective population size

Even though an actual population may be very large, its effective population size could be small. Various factors affect effective population size. First, variation in offspring number reduces N_e, and the effective population size equals approximately the actual population size divided by variance of offspring number. Second, inbreeding decreases effective size, and the inbreeding effective size is equal to the actual population size divided by $(1 + F_I)$, where F_I is the inbreeding coefficient. Third, N_e also depends on the ratio of females to males. If a population consists of N_m males and N_f females, the effective population size is

$$\frac{4N_mN_f}{N_m + N_f}$$

That is, the effective population size decreases as the sex ratio becomes more skewed. Moreover, effective population sizes for sex chromosomes and autosomes are different. Effective sizes for X chromosome and Y chromosome are $9N_mN_f4N_m + 2N_f$ and $\frac{N_f}{2}$, respectively. In addition, if the population size changes through time, the effective population size is approximately the harmonic mean of the actual sizes. Finally, in the case of subdivided populations, the effective size of the ensemble of populations could be (but not always, depending on migration situation) greater than the actual size because the probability of coalescence between individuals from different subpopulations is smaller than that in an ideal population (for more details see [47]).

2.5.2 The strength of selection relative to random genetic drift

Effective population size is a key parameter that determines the rate of evolution. The relative importance of selection and random genetic drift depends on the absolute value of the product of effective population size N_e and selection coefficient s. If $|N_e s| \gg 1$, selection dominates over genetic drift and beneficial mutations are more likely to become fixed in the population, while deleterious mutations are efficiently eliminated. If $|N_e s| \ll 1$, selection is swamped and the frequency of non-neutral mutations fluctuates like neutral mutations. If $|N_e s| \approx 1$, selection and genetic drift both play important roles in determining the evolutionary fate of a new mutant allele. These principles not only imply that the efficacy of selection increases with N_e but they also imply that more genes are affected by selection in populations with a larger effective population size. Mutations with the same selection coefficient therefore can have different fates in populations with different effective sizes. Factors that reduce effective size, such as large variance in offspring number, inbreeding, and skewed sex ratio, reduce the efficacy of selection.

2.5.3 How do we estimate effective size from sequence data?

Effective population size can be directly estimated by genetic diversity if the mutation rate of the organism is known. Under neutrality, genetic diversity is proportional to the product of the effective population size and the mutation rate ($\pi = 4N_e\mu$ and $\pi = 2N_e\mu$ for diploid and haploid systems, respectively). Alternatively, if temporal polymorphism data are available, variance effective population sizes can be estimated by variance in allele frequency changes. If there are large changes in population size, these two methods could give very different estimates. The first N_e represents the effective population size over longer time periods, whereas the second N_e is more related to current population size. If we want to predict the effectiveness of selection in the near future, the current N_e is more suitable, but if inference of past evolutionary processes is the goal, then the long-term N_e is appropriate.

As an example, we may consider the effective population size of *P. falciparum*. Based on the DNA sequence diversity of nuclear and mitochondrial genes, N_e has been estimated to be on the order of 10^5 [52, 77] , an order of magnitude larger than the effective population size of humans. It would be interesting to compare this estimate with that of N_e based on changes in allele frequency across generations in contemporary populations as the latter

could be considerably smaller. If so, this would mean that selection for drug resistance or virulence may become less effective as the effective population size decreases. The finding would also have implications for increased contributions of random genetic drift in changing allele frequencies in parasite populations.

2.6 Variation in mutation rates

In addition to effective population size, mutation rate is an important parameter in evolutionary processes. Mutation rates are important because they influence the speed of adaptation to new environments or selection pressures, the magnitude of mutation load due to deleterious mutations, and other evolutionary dynamics, such as coevolution between host immune systems and pathogenic strains. For neutral mutations, Kimura (1968)[53] showed that the evolutionary rate of neutral substitutions equals the rate of neutral mutations. This important principle is surprisingly simple to derive: The rate of nucleotide substitution is equal to the product of the probability of fixation and the mutation rate. For newly arisen neutral mutations in a diploid population, the fixation rate is $1/(2N_e)$. The total mutation rate equals $2N_e$ times the neutral mutation rate per individual, say, U. Therefore, the neutral substitution rate is expected to be $\frac{1}{(2N_e)} \times 2N_e \times U = U$.

Many studies suggest that, rather than being uniform across the genome of an organism, mutation rates vary across the genome. Wolfe et al. (1989)[121] found that the rate of synonymous substitution, for which it is widely assumed that $|N_e s| \ll 1$, is not uniform among genes in the mammalian genome. Studies of mammalian repetitive sequence showed a similar variation. Furthermore, Gaffney and Keightley (2005)[33] compared substitutions in repetitive elements in mice and suggested that mutation rates vary on a megabase length scale in the murid genome. On the other hand, Amos [4] observed small clusters of SNPs in the human genome.

Several confounding factors might affect inference of variation in mutation rate based on sequence data. First, natural selection shapes local nucleotide variation. Genes under positive selection tend to have lower intraspecies polymorphism and higher interspecies divergence, whereas genes under purifying selection are more likely to have lower polymorphism and lower divergence than neutral loci. Second, recombination breaks down associations between linked alleles and makes the history of different regions in the genome different.

Polymorphism would vary among different segments that have different times to the most recent common ancestor because of recombination.

To avoid the confounding effects of sequence-based methodology and to achieve more accurate estimates of mutation rate, Lang and Murray [60] improved the method of estimating mutation rates based on the fluctuation test in yeast and applied this experimental method to the estimation of mutation rates of two genes. The mutation rates were estimated on a per-nucleotide scale and were shown to differ significantly, supporting the view that the mutation rate is not uniform over the yeast genome.

Other researchers have investigated factors that may affect mutation rates across the genome. First, mutation rate is associated with DNA replication timing in humans and is higher in later-replicating regions of the human genome [104]. Second, Amos [4] suggested that mutations tend to occur near existing polymorphic sites. This author conducted computer simulations of SNPs and compared the simulated pattern with actual SNPs of human chromosome 1 and found that a nonindependence model predicted the frequency and density of the SNP clusters better than the random model. This result suggests that mutations are more likely to occur in regions that are already polymorphic. This mechanism is attractive since it seems to be favored by selection: Polymorphic regions are usually able to tolerate or even favor mutations, and fewer mutations in nonpolymorphic regions can reduce the potential deleterious mutations. This idea will be more convincing if the proposed mechanism can be supported by more direct experimental evidence. If more mutations take place near sites with existing mutations, it also implies that mutations are more likely to happen in genes under balancing and diversifying selection and further suggests that the speed of evolution of antigenic genes could be increased by both mutation and selection.

Variation of mutation rate across the genome reduces effectiveness to identify putatively functional regions. Researchers have been trying to reduce this effect: In coding regions, synonymous changes are usually used as a control for mutation rates by assuming that synonymous changes are effectively neutral. If the numbers of synonymous changes and nonsynonymous changes are both high or are both low, it is likely that they are affected by higher or lower mutation rates in the region. If there are many nonsynonymous changes but few synonymous changes within a gene, the large number of nonsynonymous changes cannot be explained by mutation rates and suggests that positive selection may be acting on this gene. However, these methods are problematic. For example, synonymous sites are

not always nearly neutral, and besides non-neutral synonymous sites within coding regions, increasing amounts of noncoding DNA have been found to be functional, making it important to annotate these functional regions in noncoding DNA. To improve the efficiency of identification of putatively functional regions, further work is needed to estimate mutation rates on a fine scale across the genomes.

2.7 What can we learn from polymorphism and divergence?

Advances in sequencing technology and other aspects of genomics have enhanced the power to study basic evolutionary processes and their relationships to diversity. Sequence data can be used to infer the demographic history of a population, the population structure, and which genes are under selection. It can be used to trace the evolution of antigenic genes in order to understand the evolution of virulence and drug resistance in microbes. In this section, we briefly introduce some methods that are useful in the study of evolution of microbes and we describe what kinds of questions can be answered.

Linkage disequilibrium would appear if individuals in the population tended to mate non-randomly due to geographic distance or had mating preference with particular phenotypes. Natural selection can also lead to linkage disequilibrium by favoring specific combinations of alleles at different loci. Therefore, it is important to examine the possibility of population structure before identifying genes under selection. The difference between the effects of natural selection and population structure is that natural selection acts only on some genes, whereas population structure affects the pattern in the whole genome. There are two kinds of approaches for detecting population substructure. One is based on clustering methods, such as those implemented in software packages *Structure*[95] , FRAPPE [107] , SABER [108] , and ADMIXTURE [2] ; the other approach is principal component analysis (PCA), implemented in the SmartPCA application [92].

When population size increases, more new mutations occur, increasing the number of low frequency alleles in the population. Conversely, a reduction in the population size results in a deficit of rare alleles. Since changes of population size affect the relative amount of alleles with different frequencies, an allele frequency spectrum (distribution of allele frequency) can be used to examine whether population size has changed in the recent past. There are some software packages available such as BEAST and $\delta a \delta i$ that can estimate demographic

parameters. The method implemented in BEAST is based on Bayesian analysis, whereas $\delta a \delta i$ applies a diffusion approximation to the allele frequency spectrum for demographic inference.

Selection forces can also affect the allele frequency spectrum, but, similar to population structure, any change in population size affects the whole genome, while selection forces only influence a limited number of genes. Any pattern pervasive across the genome is likely caused by demographic changes.

As previously discussed, effective population size can be estimated from genetic diversity if the mutation rate is known or from allele frequency changes if temporal allele frequency data are available. Moment methods, maximum likelihood methods, and Bayesian approaches based on the temporal allele frequency data are all available, and the software package NeEstimator includes many of these.

2.7.1 Identifying genes putatively under selection

Identifying genes under selection has always been a major interest in population genetics. Identifying genes under selection is important in understanding interactions between pathogens and their hosts, as genes related to host parasite interactions are expected to be subject to natural selection. Inference of selection may also be useful for choosing targets of new therapeutics. To identify genes under selection, investigators typically compare divergence between species, polymorphism within species, or both.

Genes under positive selection are expected to have higher divergence among species and genes under negative selection to be less diverged. In order to control for different mutation rates, the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site (dN/dS) is used as an indicator of selection. The software package PAML can be used to calculate dN/dS and to test for statistical significance. $dN/dS > 1$ indicates positive selection; $dN/dS < 1$ represents purifying selection; and $dN/dS = 1$ suggests that the gene is not under selection.

Three approaches are commonly used to detect selection based on data for intraspecies polymorphism. One set of methods is based on the allele frequency spectrum, one set on linkage disequilibrium, and another set on allele frequency differentiation. How different

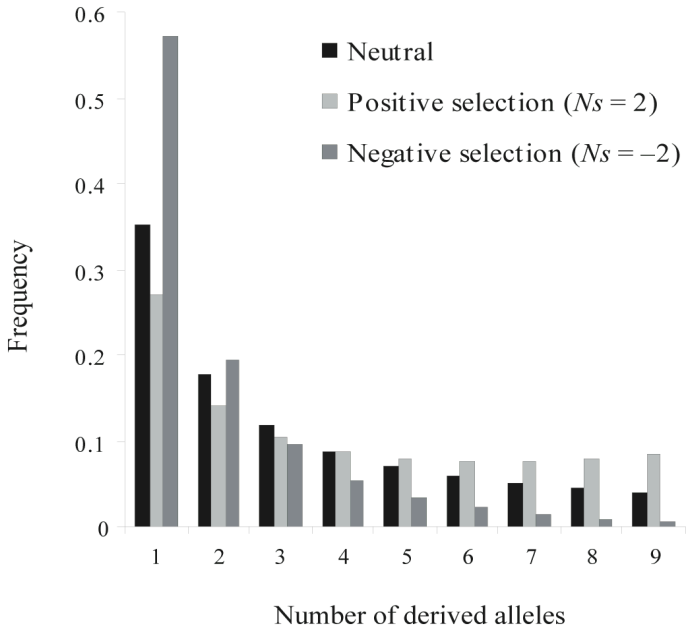


Figure 2.7.1: Allele frequency spectra. Allele frequency spectra under neutrality, positive selection, and negative selection. Equation 11 from Sawyer and Hartl (1992) was used to generate the spectra under positive and negative selections. The example here has sample size equal to 10.

selective forces change the allele frequency spectrum is shown in Figure 2.7.1. Under negative selection, deleterious mutations are removed from the population and therefore the fraction of low frequency alleles is higher. Under positive selection, mutations are favored and there is an increase in fraction of high frequency alleles. Because demographic history can alter the null distribution under the neutral hypothesis, if we detect demographic changes in the past, it is better to estimate the null distribution by coalescent simulation rather than to assume a constant population size under the neutral hypothesis. Coalescent simulation of simple demographic models can be done in a program denoted *ms* [49] .Tests for selection based on the allele frequency spectrum include Tajima's D [106], Fu and Li's D [32] , and Fay and Wu's H [29].

If positive selection is recent, recombination does not have enough time to break down linkage disequilibrium near the selected position, so linkage disequilibrium is higher and haplotype length longer around the selected loci. Methods based on haplotype length and linkage disequilibrium include integrated haplotype score (iHS) [112] and cross population extended haplotype homozygosity (XP-EHH) tests [101]. An estimation of recombination

rates is needed for applying this kind of method. Recombination rates can be estimated by experimental crosses or from linkage dis-equilibrium. Programs such as LDhat and PHASE can be used for estimating recombination rates.

Selection changes allele frequency differentiation among populations depending on the type of selection. Regions with local adaptations are more differentiated among populations, whereas regions under uniform balancing selection are less diverged. Allele frequency differentiation among populations is usually summarized by the statistic

$$F_{ST} = \frac{(H_T - H_S)}{H_T},$$

where H_S is the average heterozygosity in subpopulations and H_T is the average heterozygosity in the whole population assuming Hardy-Weinberg equilibrium [115]. Since population differentiation could also affect F_{ST}, a null distribution under a proper model of population subdivision should be generated in order to detect loci under selection by conducting F_{ST}-based method. In haploid organisms, H_S and H_T are "virtual" heterozygosities calculated from the observed allele frequencies under the assumption of random mating.

Recently, Grossman et al. [40] developed a method called composite of multiple signals (CMS) to incorporate different kinds of information, including linkage disequilibrium, allele frequency spectrum, and allele frequency differentiation among populations, to increase the power of detecting recent positive selection. This method successfully increases the resolution of localizing regions under recent positive selection by up to 100-fold.

The widely used McDonald-Kreitman method is based on the number of non-synonymous and synonymous mutations within and between species and the assumption that synonymous changes are neutral [71]. This test is more robust to the demography since synonymous and nonsynonymous mutations are similarly affected. Under a neutral prediction, the ratio of the number of nonsynonymous polymorphisms to the number of synonymous polymorphisms should be the same as the ratio of the number of nonsynonymous fixations to the number of synonymous fixations. Positive selection is expected to lead to an increased ratio of nonsynonymous-to-synonymous fixations but to have a smaller effect on nonsynonymous-to-synonymous polymorphism; negative selection is expected to result in a reduced ratio of nonsynonymous- to-synonymous fixations but to have a relatively smaller effect on nonsynonymous- to-synonymous polymorphism.

2.8 Conclusions

With the introduction of next-generation sequencing technologies, obtaining sequence data is no longer a limiting factor for large-scale population-level research. Furthermore, third-generation sequencing methods based on semiconductors are emerging. Although these new technologies offer huge advantages and massive amounts of new data, they also bring new tasks. With scale-up of the quantity of data comes a similar scale-up in the potential errors that can be introduced, such as sequencing errors, alignment errors, and missing data. Sequence reads generated by next-generation sequencing methods are short, and thus the difficulty of alignment is challenging. This can be partially solved by using paired-end sequencing, but the alignment for repetitive or highly polymorphic regions is still difficult, and correcting for biases in estimating nucleotide diversity still needs to be solved. Most sequencing errors are visible as singleton alleles observed once per sample and skew the allele frequency spectrum toward singletons. Since some of the methods for inferring demographic history and detecting selection are based on allele frequency spectrum, they are sensitive to sequencing error. Missing data are also a big issue that remains to be solved. Simply ignoring missing data can bias the population genetic inference because often data are not missing at random, but imputation can introduce different biases. When using next-generation sequencing data to study population genetic questions, researchers need to be aware of the biases that sequencing error and missing data introduce. One way to reduce the effect of sequencing error is to incorporate quality scores in the analyses. Rather than using all the sequence information that passes the quality score threshold to make inferences, we can include a quality score that is related to the error probabilities in a statistical model. Lynch [64, 65] developed a method to incorporate a quality score to estimate nucleotide diversity and the allele frequency spectrum. Further work to develop tools to account for sequencing errors and missing data remains to be done.

3

Rapid, field-deployable method for genotyping and discovery of drug-resistance single nucleotide polymorphisms in *Plasmodium falciparum*

Despite efforts to reduce malaria morbidity and mortality, drug-resistant parasites continue to evade control strategies. Recently, emphasis has shifted away from control and toward regional elimination and global eradication of malaria. Such a campaign requires tools to monitor genetic changes in the parasite that could compromise the effectiveness of anti-malarial drugs and undermine eradication programs. These tools must be fast, sensitive, unambiguous, and cost-effective to offer real-time reports of parasite drug susceptibility status across the globe. We have developed and validated a set of genotyping assays using high-resolution melting (HRM) analysis to detect molecular biomarkers associated with drug

resistance across six genes in *Plasmodium falciparum*. We improved on existing technical approaches by developing refinements and extensions of HRM, including the use of blocked probes (LunaProbes) and the mutant allele amplification bias (MAAB) technique. To validate the sensitivity and accuracy of our assays, we compared our findings to sequencing results in both culture-adapted lines and clinical isolates from Senegal. We demonstrate that our assays (i) identify both known and novel polymorphisms, (ii) detect multiple genotypes indicative of mixed infections, and (iii) distinguish between variants when multiple copies of a locus are present. These rapid and inexpensive assays can track drug resistance and detect emerging mutations in targeted genetic loci in *P. falciparum*. They provide tools for monitoring molecular changes associated with changes in drug response across populations and for determining whether parasites present after drug treatment are the result of recrudescence or reinfection in clinical settings.

The Malaria Eradication Research Agenda (malERA) initiative recently reported that development of new tools for surveillance and detection of drug-resistant parasites is imperative for any successful eradication effort. Rapidly expanding genetic data sets for malaria can be leveraged to identify molecular biomarkers related to important parasite characteristics, which could be key for surveillance [69].

Arguably the most well studied and clinically important sets of molecular biomarkers are genetic loci associated with resistance to antimalarial compounds. Resistance to aminoquinolines, such as chloroquine, and antifolates, such as pyrimethamine and sulfadoxine, is widespread, and molecular markers associated with these resistant phenotypes have been identified. Specifically, researchers have identified genetic changes in the *pfcrt* locus for chloroquine resistance [30], the *dhfr* locus for pyrimethamine and proguanil resistance [94, 97], and the *dhps* locus for sulfadoxine resistance [14]. Other candidate molecular biomarkers for drug resistance are *pfmdr1* [31, 118], associated with resistance to amodiaquine and other aminoquinolines; *cytB*, associated with atovaquone resistance [38, 57]; and *PfATPase6*, implicated in artemisinin resistance [19, 50, 110]. New discovery efforts, including genome-wide association studies, have identified additional candidate loci related to drug resistance [78, 111, 127]; these loci will need to be validated in independent parasite populations.

Because the parasite population changes readily and radically in response to selective pressure from drugs and other intervention methods, eradication policies must be equally responsive and adaptable. The data from these assays must allow real-time observation of

parasite population changes to inform timely protocol changes and further global eradication efforts to reduce or even eliminate child deaths from malaria.

While a number of assays have been developed to genotype drug resistance loci, they rely on technologies that are not suited to real-time detection and discovery of single-nucleotide polymorphisms (SNPs) associated with reduced drug sensitivity. Many of these technologies, like pyrosequencing and real-time PCR [120, 128], require extensive training and the use of expensive reagents on instruments in regional clinic settings, which means that samples collected from field sites can wait months or years for analysis. Other technologies, such as restriction fragment length polymorphism (RFLP) and *msp*-1, *msp*-2, and *glurp* marker typing, are economical to implement, but interpretation of results can be complicated and dependent on the user's skill level or lead to results that cannot be directly compared between studies [80]. These limitations reduce the utility of existing methodologies for global surveillance campaigns and emphasize the need for solutions that are portable, cost-effective, and uncomplicated.

In this report, we describe a novel modified high-resolution melting (HRM) method for real-time genotyping of malaria, using parasites derived from culture-adapted isolates. We validate these highly sensitive and robust assays by examining known drug resistance loci found in patient blood samples from Senegal. We also report the discovery of new genetic variants for several genetic loci known to be involved in modulating drug resistance.

3.1 Materials and methods

3.1.1 DNA extraction and quantification

We utilized culture-adapted parasites and clinical isolates from a number of sources. Culture-adapted *Plasmodium falciparum* parasites obtained from the Malaria Research and Reference Reagent Repository (http://MR4.org) included 3D7 (MRA- 151, deposited by D. Walliker), 7G8 (MRA-152, deposited by D. Walliker), Dd2 (MRA-156, deposited by T. E. Wellems), D10 (MRA-201, deposited by Y. Yu), FCR3 (MRA-731, deposited by W. Trager), V1/S (MRA-176, deposited by D. E. Kyle), W2 (MRA-157, deposited by D. E. Kyle), W2mef (MRA-615, deposited by A. F. Cowman), K1 (MRA-159, deposited by D. E. Kyle), and TM90C6B (MRA-205, deposited by D. E. Kyle). Culture-adapted *P. falciparum* parasites

obtained from the Walter Reed Army Institute of Research (WRAIR) included W2, W2mef, and WMCI. We obtained an additional 27 *P. falciparum* clinical samples under human subject informed consent conditions from patients being evaluated at the SLAP clinic in Thiès, Senegal, during 2007. The research and ethics study was approved by the Harvard School of Public Health Institutional Review Board (IRB) and the Ministry of Health in Senegal. Written informed consent was obtained in the subjects' own language.

We isolated genomic DNA from culture-adapted parasites using the QIAmp DNA Blood Mini Kit (Qiagen catalog number 51106), Qiagen G-100 (catalog number 13343), and Promega Maxwell 16 Blood DNA Purification Kit (Promega catalog number AS1010). We preserved DNA from patient blood samples by using Whatman FTA filter paper (Whatman catalog number WB120205), and isolated DNA using the QIAmp DNA Blood Mini Kit (Qiagen catalog number 51106), the GenSolve Blood Spot Kit (GenVault catalog number GVR-50), or the Promega Maxwell DNA IQ Casework Sample Kit (Promega catalog number AS1210). We quantified DNA using a Nanodrop 1000 (Thermo Scientific) or a real-time quantitative PCR assay with a labeled probe for hypothetical *P. falciparum* gene PF07_0076, as previously described [24]. We used 0.01 ng of *P. falciparum* patient sample template material per reaction.

3.1.2 Primer and probe design

We imported genomic sequence of the laboratory strain 3D7 from PlasmoDB version 6.3 (http://www.plasmodb.org) into BioFire Diagnostics, Inc. Primer Design Software (version 1.0.R.84). We selected primer and probe sets ranking in the top 10 out of 1,000 potential set designs for SNPs located in six genes across the *P. falciparum* genome: *pfcrt* (C72, M74, N75, K76, H97, A220, N326, and I356), *pfdhfr* (N51, C59, I164, and S108), *pfdhps* (S436, G437, K540, A581, and A613), *cytB* (Y268), *pfATPase 6* (L263, E431, A623, and S769), and *pfmdr1* (N86, Y184, R371, S1034, N1042, and D1246). We designed the probes to contain the SNP of interest. Where several SNPs existed in close proximity on the genome, we designed the probe to cover as many of the SNPs as possible. We determined likely melting peak separation between mutant and wild-type alleles on these probes using Hybridization Probe Melting Temperature Software v1.5 and uMelt (BioFire Diagnostics, Inc.) [26]. These applications allowed us to determine peak separation, but the actual melting peak temperatures depended on the actual amplification reaction conditions. We discarded designs that did not show at

least 1.5°C difference between melting peaks and redesigned many to increase that difference. We ordered probes with a 3' block (a C3 spacer phosphoramidite attached to a standard 3' CpG) to prevent extension during PCR amplification. Table 3.1.1 lists primer and probe sequences.

3.1.3 Amplification conditions and equipment

We optimized the assay amplification conditions on a 96-well microtiter plate system using a gradient thermocycler (Bio-Rad iCycler), followed by HRM analysis on a LightScanner-96 (BioFire Diagnostics, Inc., Salt Lake City, UT). We performed all PCR amplification reactions using 2.5μl LightScanner master mix with LCGreen Plus double-stranded DNA (dsDNA) dye (BioFire Diagnostics, Inc.) and primers and probes at various concentrations for assay optimization. We used cultured *P. falciparum* strains with previously reported genotype profiles for wild-type and mutant SNPs to test the performance of the assays in detecting both alleles. Where we did not have samples with all possible mutant SNPs for an assay, we generated constructs containing those mutated positions and used them as controls for the assays (miniGenes; Integrated DNA Technologies). We verified all products by sequencing by Genewiz (Genewiz, Inc., Cambridge, MA) on an ABI 3730xl DNA analyzer and with BigDye Terminator v3.1 chemistry using the company's sample submission requests.

Once optimized for 96-well plate systems, we transferred the technology directly to a LightScanner-384 to increase assay economy, with decreased reaction volumes (5μl) and 384-well reaction plates instead of 96.

We performed PCR using a gradient thermal cycler to determine optimal annealing temperatures for asymmetric PCR conditions, with one primer in 5:1 or 7:1 excess and the probe at 80 to 100% of the concentration of the excess primer (data not shown). Asymmetric PCR is essential to probe-based amplification, producing excess single-stranded material that binds to the blocked probe, which increases the signal needed for probe melting analysis. For assays crtH97, cytBY268, and mdrS1034, the optimal asymmetric ratio of forward to reverse primers was 5:1; for the remaining assays, the asymmetry was 1:5. The final primer concentration was 0.5 μM excess primer and 0.1 μM limiting primer. The final reaction concentration of the 3'-blocked probes was 0.4 μM.

Table 3.1.1: High Resolution Melting assay primer and probe sequences.

Gene	Assay SNP ID	Forward Primer (5'>3')	Reverse Primer (5'>3')	Probe (5'>3')
pfATPase6	L263	CATGCTGTTATAGAATCAAATAGTGAAGATACTC	AATTAATAATCCATACAGTTACACATATTACAAAAATGA	AAAATCGATTTATTTGGTCAACA-block
	E431	TTTTTTAATAAATTAAAAGATGAAGGAAATGTTGAAGC	CATTTTCTTACTATCACTAGAAAAATAATCACTATATGG	ATTGATCCTTCTTCTCCATCATCC-block
	A623	TGAATGAAAATGATAAGAATTTAAAGAATGCTAACC	ATGCTCAAATGTATTTTCTCCTATAGCTT	TACAGCTCAGGCAACAACAAAT-block
	S769	AAATGAAATTCATAATAAGATTCAAAATATGGGA	AAATCTTGTTCTAATTTATAATAATCATCTGTATTCT	CTTTGCTTATAAAAAATTAAGTAGTAAAGAT-block
pfcrt	C72/M74/N75/K76	GTAAAACGACGGCCAGTTTCTTGTCTTGGTAAATGTGCTCA	CAGGAAACAGCTATGACCGGATGTTACAAAACTATAGTTACCAAT	GTGTATGTGTAATGAATAAAATTTTTG-block
	H97	TTTGCTAAAAGAACTTTAAACAAAATTGGTAACTA	ATTTATCTTACTTTTGAATTTCCCTTTTTATTTCCA	CATACAAATAAAGTTGTGAGTTTCGGATGTTAC-block
	A220	GCTCAGGTCGTTGAAACACAAGAAGAAAATTCTATC	GCTCAGGCTGAAACAAAGTTTAAGTGTTAATATATATTAAATATTAC	GTCTTAATTAGTGCCTTAATTGTA-block
	N326	CGAGCATTTTTTAGAAAACCTTCGCATTGT	TTCATCCTTTTTATTCTTACATAGCTG	CTTCTTTGACATTTGTGATAATT-block
	I356	GCTCAGGTCGAAATTTTCTACCATGACATATACTATTG	CACCTGACTGATTTATATATTTATATCTTTTTAATTCTTACGGC	GTCCAGCAACAGCAATTGCT-block
pfcytB	Y268/M270	GTAAAACGACGGCCAGTTGATAATGCTATCGTAGTAAATA CATATGTTACTC	CTTTGTTCTGCTAATAAGAATAATAATTGTAATGA	AACATTGCATAAAATGGTAGAAAGTAC-block
pfDHFR	N51/C59	ACATTTAGAGGTCTAGGAAATAAAGGAGT	ATATTTACATCTCTTATATTTCAATTTTTCATATTTTGATTCATTCAC	AAATGTAATTCCCTAGATATGAAATATTTTTGTGCAG-block
	I164	ACAAAGTTGAAGATCTAATAGTTTTACTTGGG	CTGGAAAAAATACATCACATTCATATGTACTATTTATTCTA	AATGTTTTATTATAGGAGGTTCCG-block
	S108	CTGTGGATAATGTAAATGATATGCCTAATTCTA	GACAATATAACATTTATCCTATTGCTTAAAGGT	GGAAGAACAAGCTGGGAAAGCAT-block
pfDHPS	S436/A437	GAATGTTTGAAATGATAAATGAAGGTGCTA	CAGGAAACAGCTATGACGAAATAATTGTAATACAGGTACTACTAAATCTCT	ATCCTCTGGTCCTTTTGTTATACC-block
	K540	GTGTTGATAATGATTTAGTTGATATATTAAATGATATTAGTGC	GTTTATCCATTGTATGTGGATTTCCTCTT	TAATCCAGAAATTATAAAATTATTAAAAAAAAAAAAC-block
	A581	CTTGTATTAAATGGAATACCTCGTTATAGGA	AGTGGATACTCATCATATACATGTATATTTTGTAAG	TTGGATTAGGATTTGCGAAGAAACATGATCA-block
	A613	CTCTTACAAAATATACATGTATATGATGAGTATCCACTT	CATGTAATTTTTGTTGTGTATTTATTACAACATTTTGA	AAGATTTATTGCCCATTGCATGA-block
pfMDR	N86	TTATTATTTATATCATTTGTATGTGCTGTATTATCAGG	CAGGAAACAGCTATGACATCATTGATAATATAAATTGTACTAAACCTATAGATACT	GAACATGAATTTAGGTGATGATATTAATCC-block
	Y184	AGTTCAGGAATTGGTACGAAATTTATAACA	ACGCAAGTAATACATAAAGTCAAACG	CCTTTTTAGGTTTATATATTTGGTCAT-block
	R371	ATGGTGCCTCAGTTATATCCATTT	TTTGTTGCTTCTAAAGCTTTCATATATTCTGT	GGTGTACTTATTAGTATGTTTATGTTAAC-block
	S1034	AAATAAAGGACAAAAAAGAAGAATTATTGTAAATGC	AGAAGGATCCAAACCAATAGGC	AGCGCTTTGACTGAATCCCCA-block
	N1042	AAGAATTATTGTAAATGCAGCTTTATGGG	GGATTTCATAAAGTCATCAACTAATATAGTACCTC	CAATTATTTATTAATAGTTTTGCCTAT-block
	D1246	GCAGAAGATTATACTGTATTTAATAATAATGGAGA	TTTCATATATGGACATATTAAATAACATGGGT	GTGATTATAACTTAAGAGATCTTAGAAACT-block

The best amplification and melting conditions depended on the instrument used for HRM. On the LightScanner-96 or -384, 55 cycles of a three-step program performed best. For assays dhfrS108, mdrD1246, ATPase6S769, crtI356, and crtA220, the reaction mixtures were denatured for 2 min at 95°C and denatured, annealed, and extended at 94°C, 63°C, and 74°C for 30 s each, with a final step at 95°C for 30 s. All remaining assays had an annealing temperature of 66°C instead.

With glass capillaries used in the LightScanner-32 (LS32; BioFire Diagnostics, Inc.), which transfer heat much more efficiently than the plate- based LightScanner-96 and -384, the amplification cycle was 95°C denaturation for 2 min, followed by 55 cycles of 94°C for 5 s and 66°C for 30 s, and then a premelt cycle of 5 s each at 95°C and 37°C. Following PCR, the amplified products were analyzed using HRM on the LightScanner instruments. Products were heated from 40 to 90°C and the change in fluorescence recorded incrementally. Standard software included with the instruments were used for unlabeled probe analysis to visualize melting peaks based on these changes in fluorescence.

Once we selected an optimal annealing temperature for pure allelic samples, we implemented a novel amplification protocol to preferentially increase the amplification of mutant alleles in mixed samples. The mutant allele amplification bias (MAAB) technique is an experimental approach used in samples suspected of containing very small fractions of mutant alleles in a single sample. MAAB uses decreased annealing temperature during PCR amplification to increase the melting signal from the mutant allele in mixtures. Amplification bias of the mutant allele is accomplished by using a probe design that is perfectly matched to the normal, or wild- type, allele.

We developed MAAB protocols on a LightScanner-32 to take advantage of the high temperature transition rate capability, which is critical for optimal performance of the MAAB technique. This instrument utilizes capillary tubes in a 5- to 10-μl reaction volume and offers real-time PCR and automated high-resolution melting analysis capability. The glass capillary reaction vessels accommodate the requisite rapid temperature transitions required for MAAB. Optimal annealing temperatures under MAAB were reduced to 52 to 56°C.

3.2 Results

We designed a set of 23 HRM assays using the 3'-blocked LunaProbe technology to genotype SNPs across six genes previously associated with reduced drug sensitivity. These assays failed traditional TaqMan assay design, but HRM design was successful in all cases. The LunaProbe technology has been shown to increase assay sensitivity: the addition of 3'-blocked probes covering short (less than 50-bp) regions within the primer amplicon results in enhanced peak melting-temperature (Tm) separation between alleles, especially class IV SNPS (A→T or T→A) that are a challenge to differentiate with traditional HRM methodologies. Figure 3.2.1 A and B illustrate this using Pfcrt H97. Shown is TM90C6B with the mutation H97Q that results from a nucleotide base change from A to T. This class IV SNP is the most challenging for HRM to distinguish. Figure 3.2.1 B is an example of detection of multiple alleles. Pfmdr 86 shows wild-type N86 in 3D7 (gray), mutant N86Y in strains K1 and FCR3 (red), and a recently reported mutation in field isolates, N86F in Dd2 (blue). Note that different sources of Dd2 have varying numbers of copies of pfmdr1. In this instance, Dd2 has copies with different SNP mutations that have been sequenced to verify the presence of both alleles. Both N86Y and N86F derive from an A→T nucleotide change. Additionally, the SNPs tested are adjacent to one another in the genome and can be clearly differentiated by the assay.

LunaProbes also allow separation of SNPs located very close or adjacent to each other in the genome (Figure 3.2.1 B and C), allowing determination of SNP haplotype blocks. Figure 3.2.1 C uses Pfcrt 72-76 to illustrate. The probe is designed as a perfect match to the 3D7 allele at each locus (codons 72, 74, 75, and 76); thus, sample 3D7(blue) displays the highest melting peak (Tm) possible with this probe. This peak corresponds to a genotype at these four loci of C,M,N,K. Sample 7G8 (red) is mismatched under the probe at codons 72 (TT mismatch) and 76 (CT mismatch) for a genotype of S,M,N,T. Samples Dd2, V1/S, and FCR3 (gray) are all the same genotype and are mismatched to the probe at codons 75 (GT and AA mismatch; 1st and 3rd bases of codon 75) and 76 (CT mismatch), making the genotype C,I,E,T. If a sample were mismatched with the probe at only a single base site, the melting peak would be somewhere between the red and blue peaks. If a sample were mismatched at all four sites, the melting peak would be even lower than the gray peak.

This improved design resulted in better peak separation than previously reported with

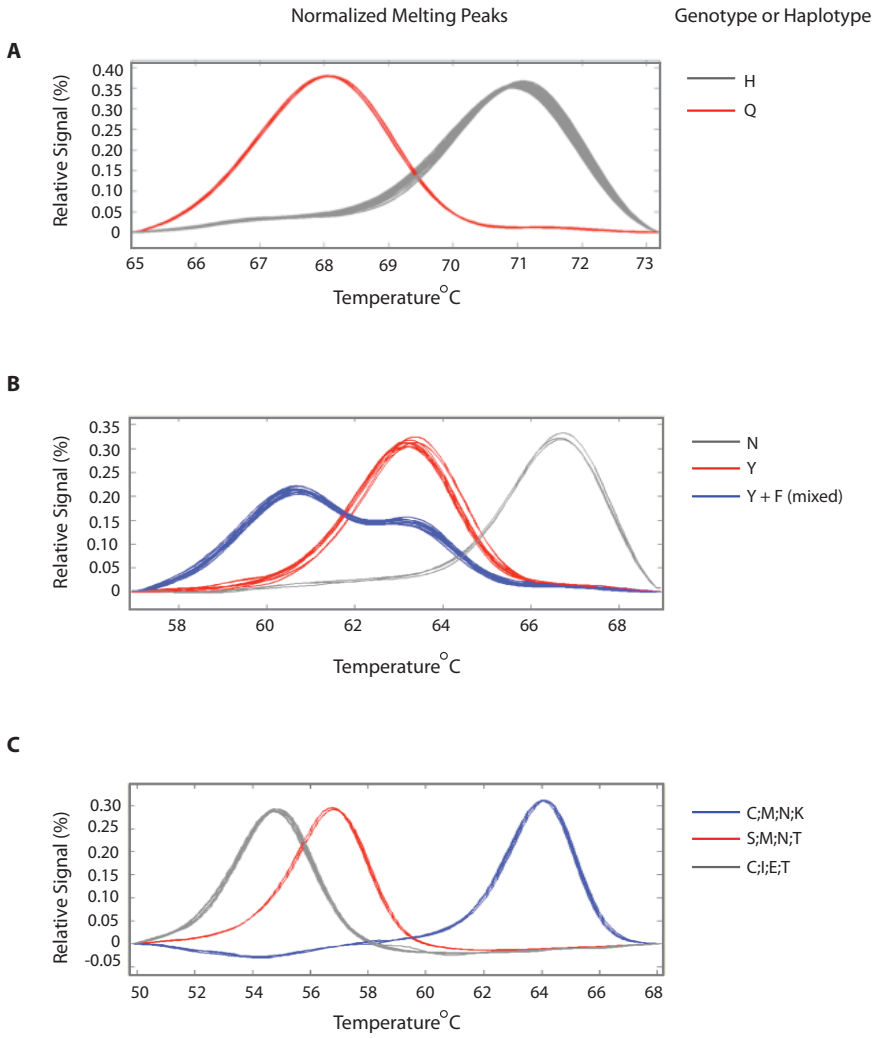


Figure 3.2.1: Representative melting peaks of High Resolution Melting (HRM) assays. (A) Example of a single-SNP probe assay; Pfcrt H97 shown. (B) Example of a single-SNP probe assay with multiple alleles in Pfmdr 86. (C) Representative SNP haplotype probe assay. Shown is pfcrt 72, 74, 75, and 76.

HRM [7]. We confirmed the genotypes determined by HRM analysis through direct sequencing of the amplicons, finding 100% correspondence of the observed genotypes to the reported sequences.

Our genotyping assays proved to be robust and specific. They were able to detect and differentiate alleles with low concentrations of template material even in the presence of

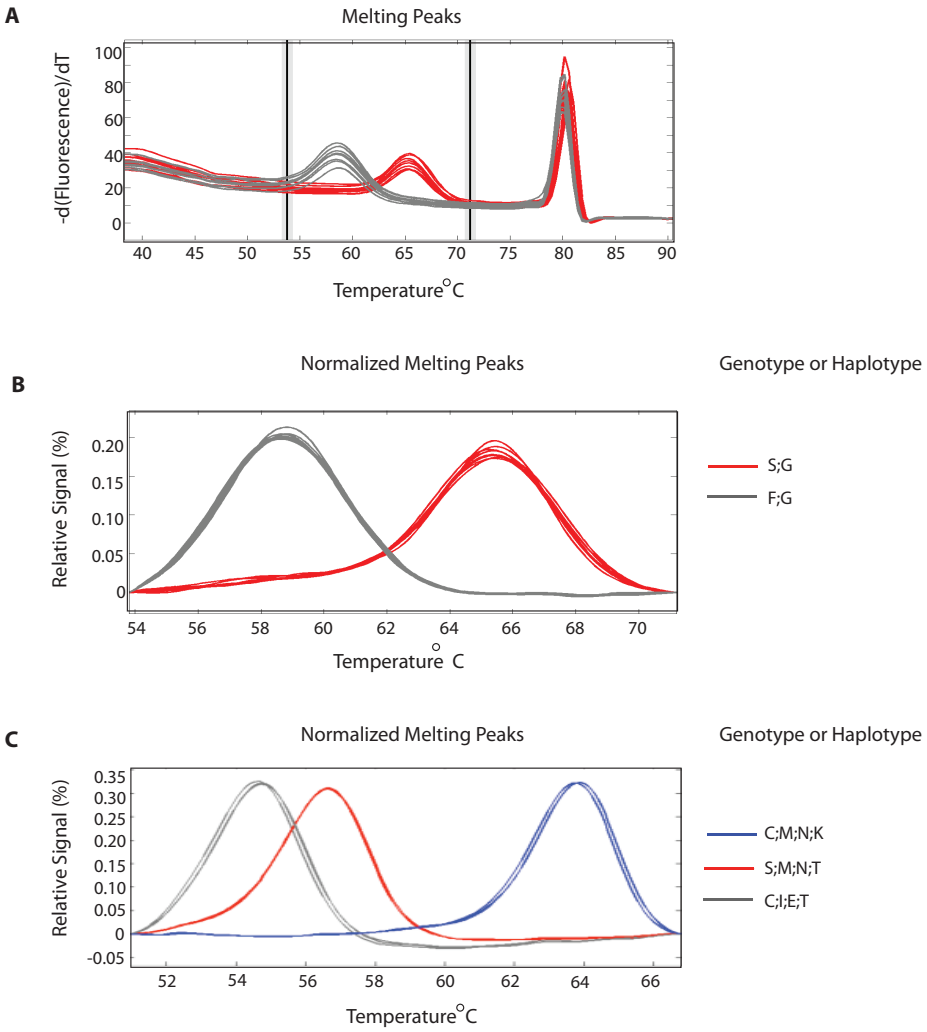


Figure 3.2.2: Limit of detection and performance with human genomic material. (A) Shown are the normalized derivative melting peaks of both the probe and full amplicon (the higher-temperature melting peaks outside the vertical bars of the probe amplicon) of representative assay (pfdhps 436/437).The melting peaks are still distinguishable, with 10^5 ng of *Plasmodium* template. (B)The same assay shown in panel A after software normalization of the probe melting peaks (BioFire Diagnostics, Inc., standard software suite analysis of unlabeled probes) still shows strong peaks for the mutant and wild-type alleles. (C) Performance with an excess of human material. Mock samples of culture-adapted parasite genomic material with an excess of human material added. Pfcrt K72-76 assay is shown. The limit of detection is 10 pg of parasite template combined with 1 ng of human DNA.

excess contaminating genomic material. Using cultured isolates without excess human material, as would be found in filter-extracted patient samples, the limit of detection was 10^5

ng, similar to that found by other groups (Figure 3.2.2 A and B) [7, 35] and also similar to other PCR-based methodologies and comfortably within the yield from extraction of patient samples collected on filter paper in the field. The assays remained specific for their target regions in the presence of an excess of human genomic material that we introduced in order to mimic the ratios of human and parasite DNAs obtained from patient samples; however, the absolute limit of detection with contaminating human DNA was reduced to approximately 10 pg (Figure 3.2.2C).

The genotyping assays are also sensitive and were made more so by our technology modifications. We were able to detect alleles comprising 2% to 5% of the total sample; by introducing the novel amplification method MAAB on a LightScanner-32 instrument using glass capillary tubes, we were able to increase the sensitivity to detect mutant alleles present at less than 1% of the mixture (Figure 3.2.3B). There is some vertical spread between the mixture peaks; however, the inflection for each mixture does not change compared to the negative or peak curve seen with the unmixed samples. This inflection difference is consistent across repeated trials and is reliable as long as unmixed controls are present.

Remarkably, we could discriminate individual alleles even in more complex mixtures containing three or more alleles using our modified HRM approach (Figure 3.2.3A). These results demonstrate that the assays are sensitive and specific and are able to identify alleles present at less than 1% of a mixture when multiple alleles exist in a sample. Comparable amplification technologies, such as TaqMan, fail to consistently detect mutant alleles at less than 10% of the sample mixture [24]. Without the refinements to HRM using blocked probes, the typical sensitivity for detection is still an improvement at 2 to 5%.

Additional methods, such as nested-PCR amplification and RFLP analysis, rely on more subjective and complicated interpretation of gel bands. Here, we show that HRM analysis results in clearly separated and unambiguous melting peaks to determine sample genotypes.

To assess the accuracy and reliability of these assays, we applied them to genotype 44 independent, culture-adapted parasites from a number of different geographical regions and verified our results by sequencing (Genewiz, Inc., South Plainfield, NJ) (Table 3.2.1). From our assays, we found that 40% of the samples had the mutant genotype for *pfcrt* K76 and 18% and 40% displayed variants at *dhps* residues 436 and 437, respectively. For *dhfr*, 65% were mutants at residues 51 and 59, and 70% were mutants at residue 108. While many

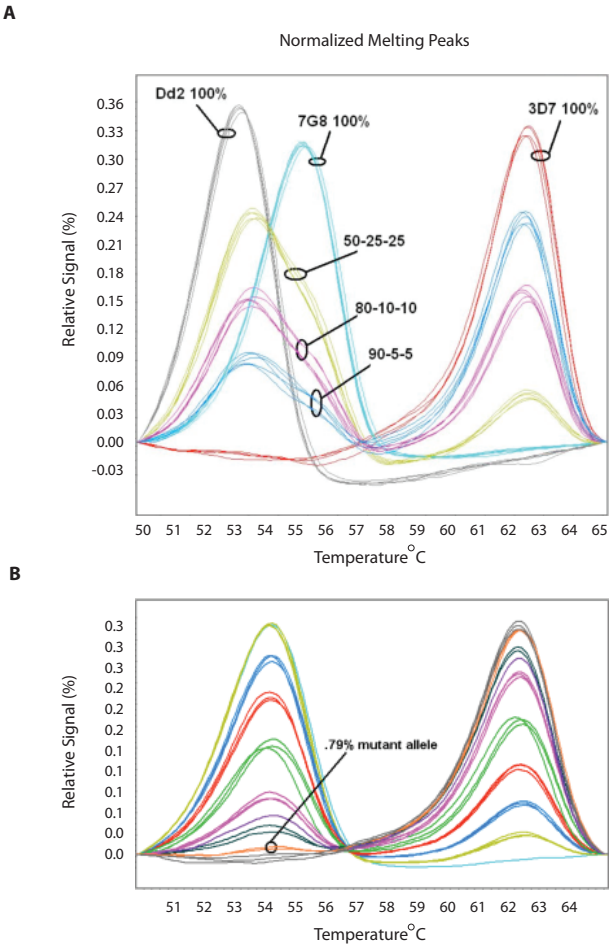


Figure 3.2.3: Assay performance with mixtures of genomes and MAAB. (A) Mixture of three genomes showing clear differentiation between fractions: 3D7-7G8-Dd2. The pfcrt K72-76 assay is shown. (B) Shown in the pfcrt K72-76 assay, MAAB in the LightScanner-32 increases the likelihood of detecting low-frequency allele fractions in a mixed population. Different proportions of mutant DNA were mixed with wild-type DNA: 100% wild-type DNA (3D7), a 50/50 mixture of mutant DNA (7G8) with wild-type DNA (3D7), a 25/75 mixture of mutant DNA (7G8) and wild-type DNA (3D7), down to less than 1% mutant DNA mixed with wild-type DNA in 1 ng total template concentration.

Table 3.2.1: Sample and genotype information.

Sample	Origin	Sample Type	Gene and SNP location																														
			pfcrt								pfmdr						pfdhps					cytB		pfATPase6				pfDHFR				
			C72	M74	N75	K76	H97	A220	N326	I356	N86	Y184	R371	S1034	N1042	D1246	S436	A437	K540	A581	A613	Y268	M270	L263	A623	E431	S769	N51	C59	I164	S108
HB3	Honduras	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	N
Dd2	Indochina	culture adaptation	C	I	E	T	H	S	S	T	F	Y	R	S	N	D	F	G	K	A	S	Y	M	L	A	E	S	I	R	I	N
MP2475	Malawi	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
MP2533	Malawi	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
PS189	Mali	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	S	Y	M	L	A	E	S	N	C	I	S
3D7	Netherlands	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	N	C	I	S
APO41	Nigeria	culture adaptation	C	I	E	T	H	S	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
N29.07	Nigeria	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	N	C	I	S
N59	Nigeria	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
N60	Nigeria	culture adaptation	C	I	E	T	H	S	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
N62	Nigeria	culture adaptation	C	I	E	T	H	S	N	I	Y	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
N92	Nigeria	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
P31.01	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
P05.02	Pikine, Senegal	culture adaptation	C	I	E	T	H	S	S	I	Y	F	R	S	N	D	A	!	K	A	A	Y	M	L	A	K	S	I	R	I	N
P11.02	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	S	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
P27.02	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	N	I	Y	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
P51.02	Pikine, Senegal	culture adaptation	C	I	E	T	H	S	S	I	Y	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
P60.02	Pikine, Senegal	culture adaptation	C	M	N	K	H	S	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
P08.04	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	A	!	K	A	A	Y	M	L	A	E	S	I	R	I	N
P09.04	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
P19.04	Pikine, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
P26.04	Pikine, Senegal	culture adaptation	C	I	E	T	H	S	N+S	I	Y	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Tg001.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	F	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Tg006.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	F	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
Tg060.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Tg103.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	Y	R	S	N	D	F	A	K	A	A	Y	I	L	A	E	S	I	R	I	N
Ts001.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
Ts002.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts007.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts009.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
Ts011.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	N
Ts012.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts013.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	N
Ts015.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts018.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	F	R	S	N	D	F	G	K	A	A	Y	M	L	A	E	S	I	C	I	N
Ts031.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts055.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	N+F	R	S	N	D	S	G	K	A	A	Y	M	L	A+E	E	S	I	R	I	N
Ts058.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N+Y	F	R	S	N	D	F	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts062.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts063.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	F	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts079.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts098.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	F	R	S	N	D	F	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts104.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts106.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
Ts107.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	Y	N+F	R	S	N	D	S	G+A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts108.07	Thies, Senegal	filter paper	C	M	N	K	H	S	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	E	E	S	I	R	I	N
Ts142.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts147.07	Thies, Senegal	filter paper	C	I	E	T	H	S	N	T	N	F	R	S	N	D	S	G+A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Ts164.07	Thies, Senegal	filter paper	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th10.04 D10	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	Y	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	N	C	I	S
Th10.04 H7	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	Y	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	K	S	N	C	I	S
Th015.04	Thies, Senegal	culture adaptation	C	I	E	T	H	S	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	S
Th026.04	Thies, Senegal	culture adaptation	C	I	E	T	H	S	N	T	Y	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th028.04	Thies, Senegal	culture adaptation	C	I	E	T	H	S	N	I	N	F	R	S	N	D	Y	!	K	A	A	Y	M	L	A	K	S	I	R	I	N
Th105.07	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	A	!	K	A	A	Y	M	L	A	E	S	N	C	I	N
Th074.08	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	K	S	I	R	I	N
Th230.08	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th231.08	Thies, Senegal	culture adaptation	C	I	E	T	H	S	N	T	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th232.08	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	Y	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th026.09	Thies, Senegal	culture adaptation	C	I	E	T	H	S	N	I	N+Y	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
Th130.09	Thies, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	G	K	A	A	Y	M	L	A	E	S	I	R	I	N
V34.04	Velingara, Senegal	culture adaptation	C	I	E	T	H	S	N	I	Y	F	R	S	N	D	F	G	K	A	A	Y	M	L	A	E	S	N	C	I	S
V35.04	Velingara, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	S
V42.05	Velingara, Senegal	culture adaptation	C	I	E	T	H	S	N	I	N	Y	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	I	R	I	N
V92.05	Velingara, Senegal	culture adaptation	C	M	N	K	H	A	N	I	N	F	R	S	N	D	S	A	K	A	A	Y	M	L	A	E	S	N	C	I	N
TM90C6B	Thailand	culture adaptation	C	I	E	T	Q	S	S	T	N	F	R	C	D	Y	S	G	K	G	A	S	M	L	A	E	S	N	R	L	N

other SNPs associated with drug resistance were primarily wild-type genotypes, there were variant polymorphisms in this international collection for each of the following loci: *pfcrt* 220 (40%) and 356 (12.5%), *pfmdr1* 86 (32.5%) and 184 (70%), and *PfATPase6* (40%). All mutations detected by genotyping assays were confirmed by sequencing (Genewiz, Inc., South Plainfield, NJ), and our assays showed 100% correlation with the sequencing results, as well as those reported on PlasmoDB version 6.3 (http://www.plasmodb.org). The genotyping assays are also sensitive and were made more so by technology improvements. These data validate the assays as able to reliably detect previously characterized mutations in the six drug resistance loci.

We next tested the correspondence of our genotyping assays with drug phenotypes among the culture-adapted parasites [8]. We evaluated the K76 locus of *pfcrt* associated with chloroquine resistance and the N51, C59, and S108 loci of dhfr associated with pyrimethamine resistance. We observed 100% correspondence between the *pfcrt* K76T mutation and chloroquine resistance among culture-adapted parasites. We also observed 100% correspondence between the N51I allele (100%) and the C59R allele (100%) in dhfr for pyrimethamine resistance. For the S108N allele, four of the parasites deviated from direct correspondence with pyrimethamine resistance. Three of the parasites that had the S108N allele tested sensitive to pyrimethamine. These parasites were wild type for the N51 and C59 alleles and had pyrimethamine resistance levels near the cutoff (2,000 nM) for resistance [9]. One parasite that had the S108 allele tested resistant to pyrimethamine. This parasite, however, was mutant for the N51I and C59R alleles and also had a level of pyrimethamine resistance that was just over the resistance cutoff level of 2,000 nM. Thus, we observed almost complete correspondence between the genotyped alleles and the expected drug responses for chloroquine and pyrimethamine among the culture-adapted parasites in this analysis, confirming that these molecular markers are useful for interpreting drug responses in *P. falciparum* parasites.

We then demonstrated the usefulness of our HRM assays on materials derived from blood collected on filter paper from patients seeking treatment for malaria at the Thiès, Senegal, clinic. Table 3.2.1 shows the prevalences of mutations in the following loci: *pfcrt* 76 (15%), *dhps* 436 (26%) and 437 (48%), and *dhfr* 51 and 59 (85%) and 108 (93%). Other SNPs with variant alleles include *pfcrt* 220 (26%) and 356 (22%), pfmdr1 86 (7%), *pfmdr1* 184 (70%), and *PfATPase6* 623 and 431 (7%). Based upon comparison to data from previous years, there has been a decrease in the prevalence of mutations associated with chloroquine

resistance and an increase in mutations associated with pyrimethamine resistance. The decrease in chloroquine resistance but increase in pyrimethamine resistance is consistent with the use of these drugs for malaria treatment in Senegal. While chloroquine use has largely been discontinued, use of pyrimethamine has continued (initially in a combination treatment with sulfadoxine and amodiaquine) and remains the mainstay treatment for intermittent preventative treatment for pregnant women (IPTp) in Senegal to date [123].

Because they generally require prior knowledge to design exact matches to every SNP sequence variants, probe-based technology approaches like TaqMan fail to amplify if novel variants appear within the amplicon region. However, because by design our assays allow us to detect both known and novel haplotypes and mutations in genetic loci associated with drug resistance. First, we detected previously unreported haplotypes in Senegal for pfdhps 436/437 in three of the culture-adapted parasite lines, including the 436A/437A haplotype for two parasites (P05.02 and Th105.07), as well as the 436Y/437G haplotype for another parasite (Th28.04) (Figure 3.2.4A). Second, we detected a new mutation in the cytB locus from a single patient sample (Figure 3.2.4B). This novel mutation, confirmed by sequencing (Genewiz, Inc., South Plainfield, NJ), is the M270I variant, located near the amino acid associated with atovaquone resistance in the *cytB* gene at Y268 [57]. Because this novel allele was detected in direct patient material preserved only on Whatman FTA cards that did not preserve whole parasites for culture, we were unable to test for any biological effects of the mutation on drug sensitivity. Because the probe-based analysis also allows study of the larger amplicon, we were able to discover SNPs 22 bp upstream of the probe region but still within the amplicon targeted by the ATPase6 431 assay in D10 (MRA-201) (data not shown). The additional utility to scan both single SNPs are well as genomic regions known to be under selection increases the likelihood that additional mutations associated with reduced drug sensitivity will be found in the regions targeted by the assay primers. While samples with variant melting peaks must be sequenced to characterize the newly-discovered mutation, HRM technologies offer the unique ability to quickly and economically scan samples for emerging variants.

In addition, the assays differentiated between copies of a genetic locus. The pfmdr1 locus is known to be present in multiple copies in some drug-resistant parasites, particularly those resistant to mefloquine [31, 119]. As part of this analysis, we observed different genotypes for the pfmdr1 locus in the Dd2 parasite and confirmed the presence of the newly reported

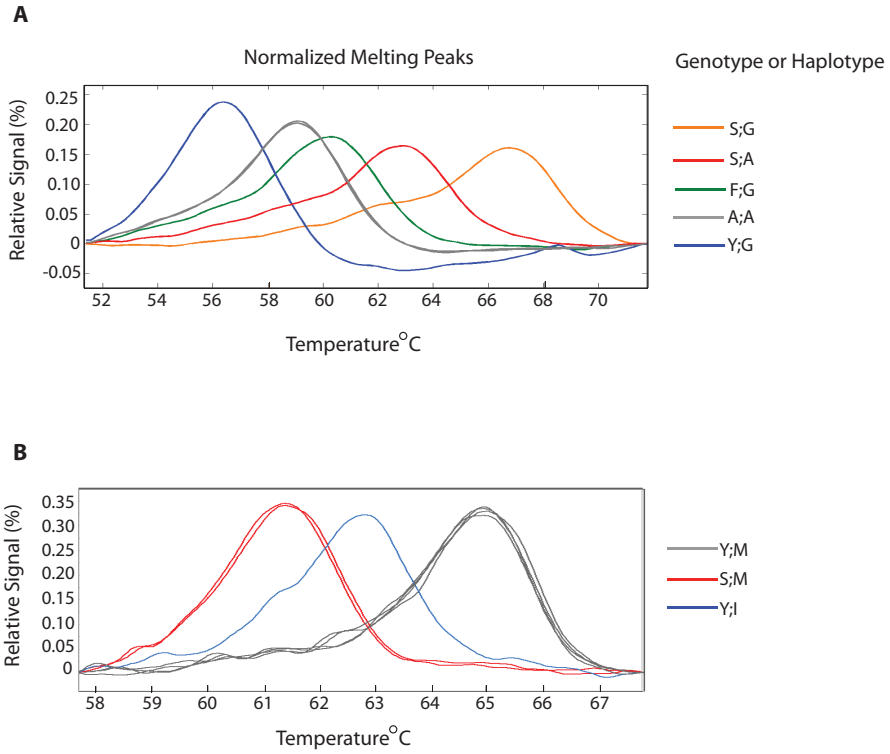


Figure 3.2.4: Detection of emerging and new mutations. (A) New SNP haplotype detected in Senegal patient samples. The pfdhps 436A/437A and 436Y/437G haplotypes have not previously been reported in Senegal, though they have been reported in other regions. (B) Novel mutation in cytB. The plot of melting peaks shows the expected wild-type Y268 and mutant S268 peaks as controls in gray and red, respectively. The new mutation, M270I, confirmed by sequencing, is shown in blue.

pfmdr1 N86F mutation in Dd2 [12, 25]. This demonstrates the ability of these assays to detect mutations among loci subject to copy number variation, which has been implicated in some drug-resistant parasites [27, 62, 82]. Copy number variation has also been identified as a source of reduced sensitivity to antimalarial drugs. Instruments with real-time capacity (BioFire Diagnostics LightScanner-32, Illumina ECO, Roche Nano and Light Cycler series, and Qiagen Rotorgene-Q) offer additional utility of these assays for direct detection of these variations.

Finally, we demonstrated that our assays distinguish between single and mixed alleles, in both culture-adapted and clinical samples. Multiple melting peaks appearing in single-genome infections determined by molecular barcode or other methods [24] suggest that there are variant copies of the genetic locus in the parasite genome (Figure 3.2.1).

3.3 Discussion

High-resolution melting technology offers a new tool for first-line surveillance of known and emerging markers associated with reduced sensitivity to drug treatments. While it does not replace standard protocols, such as sequencing, to characterize the specific genetic changes, the method is economical and easy to implement in a large number of global regions to scan the parasite population for changes that could signal the emergence or importation of known mutations. As new mutations are detected through this and other methodologies, they can be characterized by sequencing, and new HRM assays can be rapidly developed in an iterative process for population tracking.

Several other groups have presented HRM assays applied to *P. falciparum* [7, 23, 35]; however, the techniques presented here show clear advances over the previous works. We have implemented a probe-based technique that improves the sensitivity of these assays over the whole-amplicon methodology presented in the work of Andriantsoanirina et al. [7] aand similar to the technology presented by both Gan and Loh [35] and Cruz et al. [23].

Compared to the previous publications, we have developed and validated an extended set of assays for SNPs related to reduced drug sensitivity and have made notable advances in the ability to detect mutant alleles present in small amounts within a sample, an improvement on both studies, which noted a 10% limit of detection of mutant alleles while we have detected mutant alleles at less than 1%. While MAAB is most effective in instruments that use glass capillaries for their improved heat transfer properties, the method could be especially useful for surveillance of emerging resistance where the primary causative SNP is known. The ability to detect these mutations at very low levels can inform control efforts. MAAB could be applied following larger-scale plate-based screens where the results may indicate the presence of a minor population of mutants. MAAB could confirm these observations.

Further, we applied these assays to determine the SNP genotypes of 27 patient samples with both single and mixed infections and were able to detect the emergence of haplotypes new to Senegal. Like Gan and Loh [35], we identified new mutations within the probe region but also found evidence of new mutations outside the probe but within the larger amplicon.

With this technology, results are robust, consistent, and repeatable; however, care must be taken in preparing the reaction mixtures. Salt concentrations have a strong effect on the

melting temperature of the complexes, so it is advised that all samples be in the same buffer for the most consistent results. Not all samples have to be at the same concentration, but the reactions are most reliable with low concentrations of template. They become less so with an excess of template (more than 10 ng/μl), and the melting peaks may shift even if all samples are in the same buffer if any of them are at a very high template concentration relative to the other samples. While most groups report femtogram levels of detection within culture-adapted strains without human genomic material, the sensitivity of the assays decreases when this human contaminant is present in the sample. Still, we observed universally robust amplifications and unambiguous genotypes from 0.005 to 0.01 ng of *P. falciparum* template from patient-derived samples that had not been depleted of white blood cells.

HRM is a promising technology for field-based studies; however, the success of the results also depends on the quality of the analysis software that is included with the HRM analysis instruments. Currently, no standalone application exists for melting analysis. We have had the best success with the suites offered by BioFire Diagnostics, Inc., as a combination of sensitive analysis and ease of peak calling without extensive and complicated user training. User training will be further reduced as researchers publish additional validated assays that require no further optimization. Our initial panel of 23 assays is a strong start toward this goal.

Because our HRM assays are sensitive, work on mixed infections, and are easily deployed onsite to study field samples, they have the capacity to track both current and emerging infections. Our genotyping assays can be used to monitor the frequencies of known drug resistance alleles or parasite types within a population. Furthermore, they can track changes in these frequencies as interventions or other factors that impact transmission dynamics in a given population. For example, as molecular markers of emerging drug resistance are discovered, these markers can be monitored across a population to determine the likelihood that current drug therapies would remain effective or to inform policies to contain the spread of emerging drug resistance and detect the importation of drug-resistant parasites. Because these assays do not require fluorescence-based probes, the reagent cost per sample is much lower than for alternative technologies, and storage and preparation conditions are much less stringent. In our hands, HRM costs approximately 2/3 of what TaqMan-based SNP genotyping costs and have less than 1/10th the cost for sequencing all of these regions. In addition, the assays are reliable for samples collected on filter paper, a low-cost method for sample preservation, transportation, and storage. The fast readout of the melting analysis is

clear and unambiguous, allowing easy interpretation of results and less specialized researcher training.

With increased efforts to establish global networks of research centers based on-site across the globe, such as the International Centers of Excellence for Malaria Research (ICEMR) and the collaborative efforts of the Gates Foundation, the initial capital expenditures for equipment capable of high-resolution melting analysis can be distributed across sites. In addition, many of these instruments are multipurpose and are able to offer real-time analysis in addition to HRM.

Our rapid and field-deployable genotyping tools have great promise for surveillance and clinical diagnosis as we transition from malaria control to malaria elimination. For malaria elimination to be successful, tools that can provide quick and reliable information about the changing dynamics of parasites in the natural setting are key. Such information includes monitoring drug-resistant loci in parasites or determining whether a parasite persists after drug treatment. Our genotyping assays work on small amounts of patient-derived material and can disentangle information from mixed parasite infections. These assays can be used to track parasites and determine the sources of emerging infections.

4

Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal

Using parasite genotyping tools, we screened patients with mild uncomplicated malaria seeking treatment at a clinic in Thiès, Senegal, from 2006 to 2011. We identified a growing frequency of infections caused by genetically identical parasite strains, coincident with increased deployment of malaria control interventions and decreased malaria deaths. Parasite genotypes in some cases persisted clonally across dry seasons. The increase in frequency of genetically identical parasite strains corresponded with decrease in the probability of multiple infections. Further, these observations support evidence of both clonal and epidemic population structures. These data provide the first evidence of a temporal correlation between the appearance of identical parasite types and increased malaria control efforts in Africa, which here included distribution of insecticide treated nets (ITNs), use of rapid diagnostic tests

(RDTs) for malaria detection, and deployment of artemisinin combination therapy (ACT). Our results imply that genetic surveillance can be used to evaluate the effectiveness of disease control strategies and assist a rational global malaria eradication campaign.

4.1 Introduction

The *Plasmodium falciparum* malaria parasite causes nearly 700,000 deaths annually, primarily in sub-Saharan Africa [124], where disease prevalence and transmission intensity are highest. Because parasite populations are large in Africa, they are more genetically diverse there than elsewhere. They also exhibit less correlation between allelic states at different loci (i.e. less linkage disequilibrium, or LD), reflecting both the large population and also higher disease transmission rates, which facilitate sexual outcrossing [81, 111, 113].

We sought to use changes in parasite population diversity to detect longitudinal changes in disease transmission, and thereby to develop useful metrics for monitoring antimalarial interventions. As a tool to track parasite diversity, we employed a previously developed 'molecular barcode', composed of assays for 24 single nucleotide polymorphisms (SNPs) across the *P. falciparum* genome [24]. We applied the barcode to samples from Senegal. Since 2005, Senegal has dramatically increased deployment of intervention strategies, including ITNs for prevention, RDTs for detection, and ACTs for treatment, resulting in an overall decline in a number of malaria indicators [66], and making it a good site for detecting changes in parasite diversity.

4.2 Results

4.2.1 Identification of repeated barcodes

We sampled patients annually from 2006 - 2011, from the Service de Lutte Anti-Parasitaire (SLAP) clinic in Thiès, Senegal, under ethical approval, and genotyped the samples using the barcode (Methods). We first compared molecular barcodes within and between years. We confined this analysis to infections caused by a single parasite strain to reduce ambiguity from heterozygosity. The most prominent signal in our longitudinal collection of molecular

barcode data was a steady increase in the number of identical barcodes observed in distinct patient samples (Figure 4.2.1A). Whereas 10% of samples shared barcodes during the 2006 transmission season, more than 50% were within identical-barcode clusters in 2010 and 2011. Repeated instances of the same barcode were not limited to clusters of 2 or 3; in 2008 one barcode was observed in 22 distinct patient samples, and in 2011 nearly a quarter of the sampled infections exhibited another shared barcode. Overall, the proportion of unique parasite types decreased significantly over the study period (Figure 4.2.1 B; $P = 0.006$, ANOVA). We investigated whether parasite samples exhibiting identical SNP barcodes are also genetically identical at other sites in the genome by hybridizing multiple clusters of samples with shared barcodes to a whole-genome SNP array that interrogates 17,000 polymorphic positions [111]. Parasite samples sharing barcodes exhibited array-based genotype profiles as similar to each other as technical replicate hybridizations of a single laboratory reference strain (Supplementary Figure 1), suggesting that samples sharing barcodes are nearly genetically identical and likely derived from the same ancestor.

4.2.2 Clonal propagation versus epidemic expansion

The increasing occurrence of repeated barcodes (i.e. nearly genetically identical samples) in later years could be attributed to either "clonal propagation" or "epidemic expansion", or both. Clonal propagation is intrinsically linked to low parasite transmission, owing to the life history of Anopheles mosquito vectors. Female *Anopheles* mosquitoes ingest haploid *P. falciparum* gametocytes during a blood meal from a human host. The gametocytes differentiate into gametes in the mosquito midgut, where they unite to form a diploid zygote, which in turn undergoes meiosis to restore haploidy prior to inoculation of the next human host. Genetic outcrossing during the parasite s sexual stage occurs only when a mosquito bites a host infected simultaneously by multiple parasite strains and gametocytes from multiple genetically distinct strains circulate in the blood of a host; bites of singly-infected hosts result in the union of nearly genetically identical gametes in the mosquito midgut, and consequently result in self-fertilization and clonal parasite transmission. To test this possibility of increasing self-fertilization, we compared the proportion of multiple infections over time and found that the proportion of mixed infections was significantly greater in 2006-2007 compared to subsequent years (Figure 4.2.2 and Supplementary Table 1). While the patient parasitemia reported for those years varied between years, there was no trend in decreased

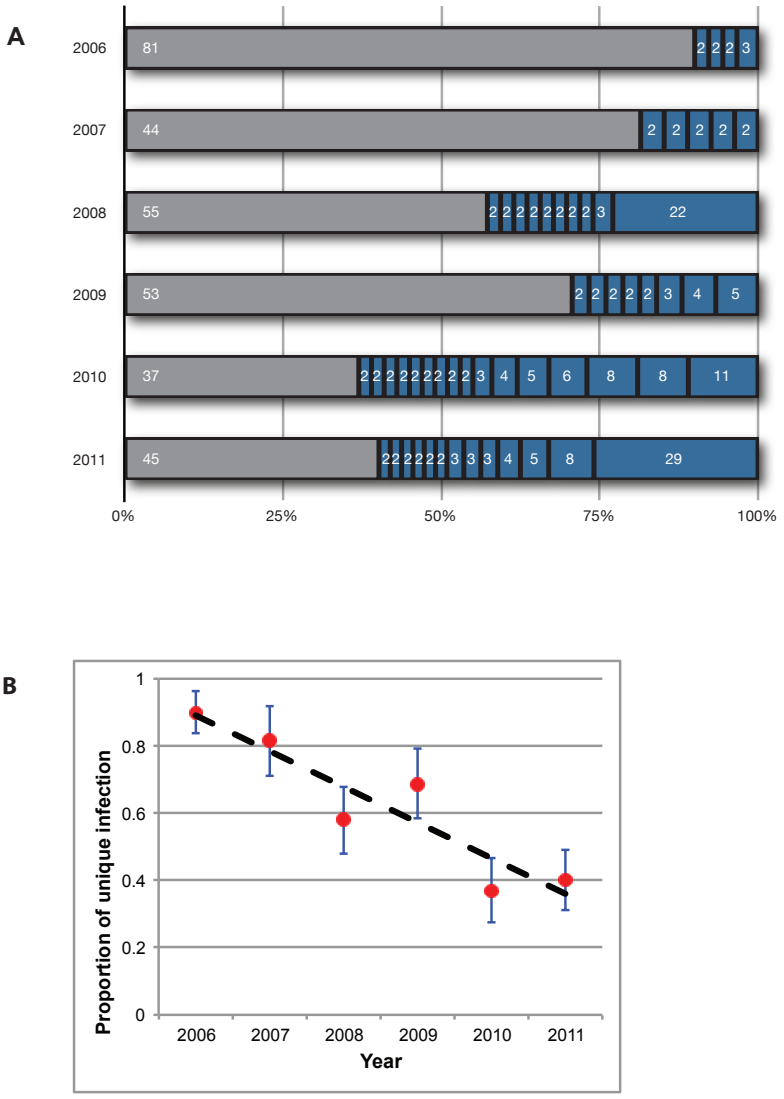


Figure 4.2.1: Temporal changes in population characteristics. A. Decreasing prevalence of unique parasite barcode profiles. For every collection season, the number of samples with unique barcodes (grey) and the number of samples in each shared-barcode cluster (blue) are shown. B. Ratio of shared vs. unique barcode profiles. The proportion of samples residing outside of shared-barcode clusters is shown per year. The error bars show 95% confidence interval of mean (±1.96 SE).

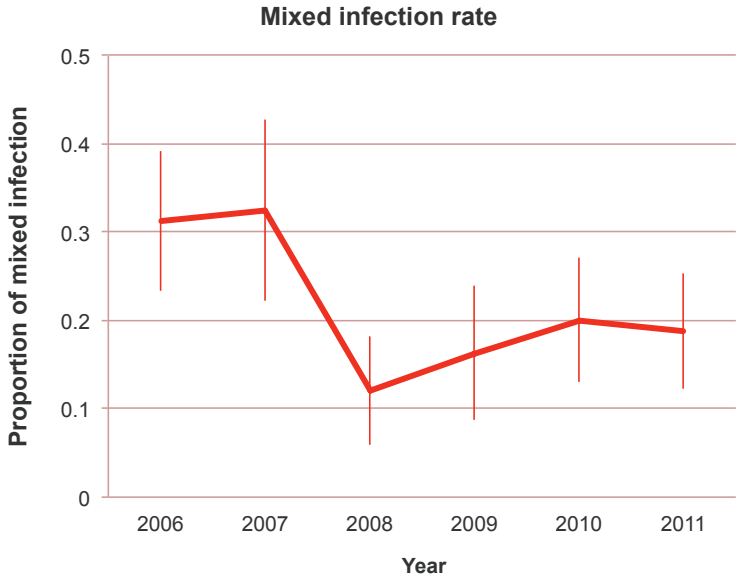


Figure 4.2.2: Mixedness over time. Proportion of mixed infections decreased between 2007 and 2008. The error bars show 95% confidence interval of mean (±1.96 SE).

parasitemia or sampling bias that could contribute to the trend (Supplementary Figure 2). This pattern of decreasing proportion of multiple infections is consistent with the decrease in the proportion of unique barcodes in Figure 4.2.1B, suggesting that clonal propagation due to decreased outcrossing is also consistent with the appearance and increase of repeated barcodes.

Epidemic expansion means that particular clones expand in the population, perhaps due to advantageous haplotypes, or a founder effect at the beginning of each transmission season, or both. Factors promoting variance in reproductive success, such as enhanced production of gametocytes, evasion of the host immune response, or enhanced transmission by selected or alternative mosquito vectors could select and enrich for favored parasite lineages in the population. Epidemic expansion is supported by the observation of two exceptionally prevalent barcodes in 2008 and 2011 (shown in Figure 4.2.3). To further test the possibility of epidemic expansion in our population, we used the framework described in Maynard Smith et al. [103] and Anderson et al.[6]. We compared multilocus linkage disequilibrium (LD) using the standardized index of association (I_A^S) [48], when including and excluding samples with the same barcode. The result shows significant LD from 2008 to 2011 when all samples

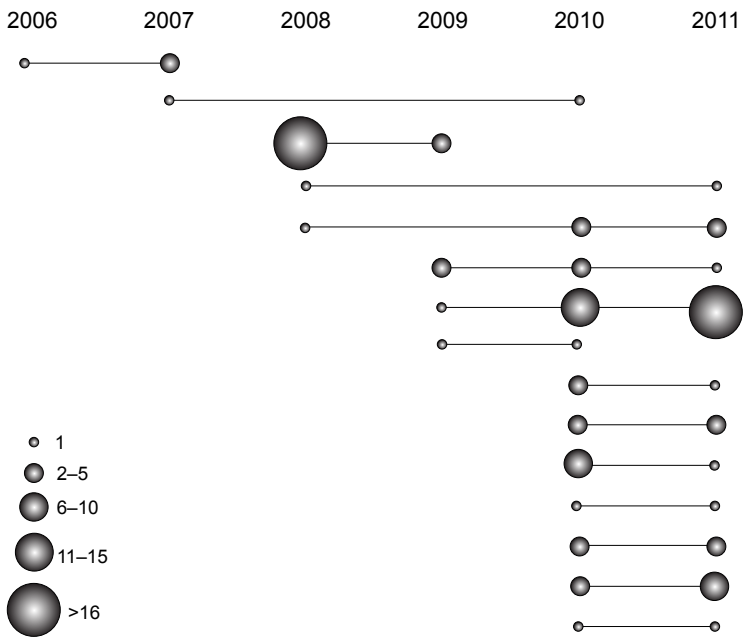


Figure 4.2.3: Clonal transmission of parasites across transmission seasons. Size and distribution of same parasite types across collection years.

are included, and no significant LD when only considering unique barcodes (Table 4.2.1), suggesting that the significant LD from 2008 to 2011 is caused by repeated barcodes; that is, some epidemic clones. There is no significant LD in 2006 and 2007 whether we included or omitted repeated barcodes. The lack of significant LD in 2006 and 2007, and the restoration of linkage equilibrium from 2008 to 2011 after excluding repeated barcodes suggest that the background population is still under linkage equilibrium and the decrease in the population recombination rate due to lowered transmission is very recent. Taken together with our analyses of the proportion of mixed infections, a likely explanation for these observations is a reduction of outcrossing in 2007-2008 followed by an expansion of individual parasite genotypes.

Moreover, we examined whether parasite samples with shared barcodes were collected in proximal dates. The difference in collection dates among samples with identical barcodes is significantly smaller than that among samples with different barcodes (Wilcoxon rank

Table 4.2.1: Multilocus linkage disequilibrium.

		2006	2007	2008	2009	2010	2011
All samples	I_A^S	-0.0049	0.0049	0.0306	0.0072	0.0293	0.1111
	P-value	0.989	0.162	$<1 \times 10^{-4}$	0.021	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$
Unique barcodes only	I_A^S	-0.0059	0.002	0.0012	0.0038	0.0042	0.0041
	P-value	0.997	0.332	0.329	0.151	0.145	0.128

sum test, $P = 0.005$), suggesting temporal expansion of particular clones in the population. However, because there was no temporal trend of increasing prevalence of a single parasite type (Figure 4.2.3) we do not believe that this was a selection event caused by emergence of drug resistance. It could possibly be a selection caused by emergence of resistance to host immune response, but the advantage disappears over time due to the corresponding changes in host, or non-selective forces. Alternatively, it might be possible that the parasite clones that appear to expand in the community were derived from an imported line novel to the area and thus the local population has little "strain-specific" immunity. We compared the pairwise differences between two exceptionally prevalent barcodes in 2008 and 2011 and the rest of strains with the pairwise differences among all strains from the same year, and found that the differences between two prevalent repeated barcodes and the rest of strains are not significantly higher than the differences among all strains from the same year (Supplementary Figure 3). This result indicates that we do not observe the evidence of imported lines from the current data. Additional sequence information of polymorphic sites will be helpful to distinguish migrants from local population.

4.2.3 Effective population size

Reduced transmission can lead to lower parasite effective population size (N_e). To test whether the deployment of intervention strategies in recent years reduces malaria transmission, we examined the parasite effective population size (N_e). Population genetic theory predicts that a decreasing population should undergo increased genetic drift, manifested as increasingly variable allele frequencies across generations. The relevant measure of effective size in this context is the variance effective population size; this we estimated by measuring the fluctuation in allele frequencies across transmission seasons of the SNPs comprising the molecular barcode. We observed large fluctuations in allele frequencies over time (Supplementary Figure 4). The variance N_e was calculated by all polymorphic SNPs using a

Table 4.2.2: Variance effective population size estimated by likelihood approximation.

	Likelihood method	
	Mean	95% Confidence Interval
2006-2007	ND	(226, ND)
2007-2008	19	(9, 49)
2008-2009	29	(12, 90)
2009-2010	18	(9, 42)
2010-2011	10	(6, 18)

* ND represents "Not Determinable".

likelihood approximation (Methods) and observed an extremely small variance N_e over time (Table 4.2.2 and Supplementary Table 2). The estimated variance effective size in 2011 is only 10, a strikingly low value that reflects large fluctuations in allele frequencies. In order to exclude the possibility that some particular parasite types are so successful in the population that they lower the estimate of effective population size, we also calculated N_e by counting each repeated barcode once (Supplementary Table 2). The estimates of N_e are still very small (less than 250) although some of the confidence intervals could not be determined. This extremely small effective population size predicts low effectiveness of selection efficiency and low rate of adaptation in Senegal.

4.2.4 Persistence across years

We also investigated the barcode dataset for evidence of clonal parasite persistence across years. Malaria transmission in Senegal is sharply seasonal, coinciding with annual rainfall patterns. Some parasite clones did indeed appear in more than one transmission season (Figure 4.2.3). These included clonal parasite types that persisted into the subsequent year and some that persisted longer, sometimes reappearing two or three seasons after initial detection. The increasing ratio of parasites persisting between years from 2006 to 2011 was statistically significant ($P = 0.008$, ANOVA) (Supplementary Figure 5). Notably, we found an increase in the frequency of identical-barcode parasites persisting between 2010 and 2011: of the 15 identical barcodes that persisted for at least one year, ten were found during that pair of years. Because parasite samples sharing the same barcode are likely to be identical by descent, the persistence of identical barcodes across years suggests multiple sequential

transmission cycles among singly-infected hosts, and indicates clonal propagation.

To explore the patterns of repeated barcodes, and to rule out sampling biases in our study design, we examined the spatial and temporal relationships between samples exhibiting identical barcodes. We insured that clonal parasites were derived from independent natural infections by assaying 18 SNPs in the human host genetic material. We found no evidence of serial sampling of the same host among samples exhibiting the same barcode (Supplementary Tables 3 and 4). Examination of patient data confirmed that barcodes observed more than once were not clustered by household, ruling out a simple hypothesis of transmission among family members. Further analysis of the parasites within these samples by sequencing of the highly-polymorphic T-epitope region of the *csp* gene provided further evidence of highly related parasites (Supplementary Table 5). We found that samples with identical barcodes are distributed across the entire transmission season and clinical catchment area, indicating a lack of temporal or spatial clustering. Our data therefore suggest a regional-level change in transmission dynamics from 2006 to 2011, rather than localized shifts.

Moreover, we compared ages of hosts before and after we observed the significant increase in the frequency of repeated barcodes. There is no significant difference in host ages between 2006-2007 and 2008-2011 (t-test, P=0.094), suggesting that the patterns of identical barcodes are unlikely to be confounded by host ages.

4.3 Conclusions and Discussion

With the restructuring of the National Malaria Control Programme (NMCP) in 2005, Senegal implemented an organized approach to malaria control and elimination. From 2006 to 2010, the NMCP increased access to insecticide-treated bednets (ITNs) and residual insecticide spraying, with the number of reported bednets per home increasing more than 35% from 2008 to 2010. Combined with no-charge access to ACTs from 2007, the country reported a 41% drop in the number of malaria cases between 2008 and 2009 [66]. The findings of increasing repeated barcodes, persistence, and proportion of single infections across transmission seasons demonstrate the usefulness of genetic tools for monitoring the effectiveness of intervention strategies against infectious disease. This type of evidence could inform control efforts as a real-time gauge of the progress towards control, elimination, or eradication. Our ability to differentiate between clonal and epidemic population structures and to track

these changes within the population could lend a more refined view of the subtle effects and varying degrees of effectiveness in control programs.

While our study reported the first evidence of clonal propagation and epidemic expansion in Africa, other groups have also used genetic tools to study parasite dynamics in geographically distinct regions, and reported clonal lineages and persistence over time [13, 39, 87]. Roper et al. showed the persistence of parasites over the dry season in Sudan and Echeverry et al. showed similar in Colombia [28, 100]. Both Branch et al. [13] and Griffing et al. [39] point to distinct genetic types within South America and Peru, in particular, and attribute population patterns to periodic epidemics in regions with relatively low transmission levels. Similarly, Nkhoma et al. [87] showed the decreases in the proportion of unique parasite genotypes and the proportion of multiple infections along with large reduction in transmission over time. However, they found no evidence of reduction in N_e during the same period of time, which was possibly caused by migrations between nearby populations, or the lack of power in analysis of temporal data when the true N_e is not small enough. Moreover, Mobegi et al. [74] showed that the background of non-clonal population structure has been widespread elsewhere surrounding our study area in West Africa, indicating that there has been dramatic changes in the population structure of this site in contrast to the surrounding regional parasite population structure. These studies, including our study, indicate the power of using genetic tools to study parasite population structure, and highlight the need for further detailed study of parasite population dynamics in more extensive geographical regions to understand the interactions and migrations between different parasite populations.

Further applications of this approach might be to differentiate between parasite recrudescence or re-emergence in selected populations to allow facile decision-making in the face of a very changeable parasite where resistance emerges quickly [24]. With additional evidence provided by other types epidemiological studies to more directly link these parameters to parasite population genetics, changes in the profile of parasites with different molecular barcodes might be used as an indicator of parasite transmission. This finding is also one beneficial outcome of a genomic diversity project undertaken by the malaria community five years ago [51, 79, 113]. The decreasing cost and increasing translation of sequencing and genotyping tools into clinical environments will make genetic data invaluable for rapidly understanding diverse aspects of infectious disease epidemiology, particularly when such information is combined with population genetic inferences and knowledge of pathogen biology.

4.4 Addendum

Following the publication of this manuscript, we have been working to address our concerns that our sampling methodology of passive case detection could be biased in representing only those *P. falciparum* patients that cause acute illness in a population already familiar with long-term and consistent exposure to the parasite. Towards that end, in 2012 we have collected samples both from symptomatic and asymptomatic patients appearing at the SLAP clinic. We determined the molecular barcode for these samples and found that the percentage of samples found within identical-barcode clusters for 2012 was nearly identical (0.53 and 0.56) for symptomatic and asymptomatic patients. This preliminary result remains to be verified with a larger number of asymptomatic patients (here N=35); however, evidence indicates that we do not have a large bias of parasite representation at this site by sampling only those patients reporting illness.

4.5 Materials and Methods

4.5.1 Study site

We obtained *P. falciparum*-positive clinical samples from patients evaluated at the SLAP clinic in Thiès, Senegal under ethical approval for human subjects and informed consent conditions. Full written consent was obtained in a protocol approved by Harvard School of Public Health, Office of Human Research Administration (P16330-110, Wirth PI) and the Ministry of Health, Senegal.

The site, located 60km southeast of the country capital of Dakar, is characterized by perennial hypo-endemic transmission with the greatest number of malaria cases by primarily *Anopheles gambiae s.l* and *A. funestus* vectors occurring approximately from September to December, at the end of the rainy season. Samples are collected passively; with patients over the age of 12 months admitted to this study with self-reported acute fevers within 24 hours of visiting the clinic and no recent anti-malarial use. Patients are screened by slide smears and rapid diagnostic test (RDT) to diagnose *P. falciparum* infection [84, 85].

4.5.2 DNA extraction and quantification

Whole blood spots from 2006-2011 were preserved on Whatman FTA filter paper (Whatman catalog #WB120205). We extracted genomic DNA from 4-6mm punches from the FTA cards using the manufacturer protocol for Promega Maxwell DNA IQ Casework Sample kit (Promega catalog #AS1210). After extraction, we quantified and generated a molecular barcode for each sample as described previously [24]. Extracted samples were excluded from analysis if the concentration (and corresponding parasitemia of the patient) were too low for successful amplification. The sample size in each year is shown in Supplementary Table 1.

4.5.3 Sequencing csp

We sequenced across the T-epitope region of the *P. falciparum* csp gene. Primer sequences were: 5 - AAATGACCCAAACCGAAATG-3 forward and 5 - TTAAGGAACAAGAAG-GATAATACCA-3 reverse. We used 1μl of each sample as a template in 25μl PCR reactions using iProof master mix (Bio-Rad cat # 172-5310) (initial denaturation 98°C 30s, followed by 35 cycles of 98°C denaturation (30s), 55°C annealing (30s), 72°C extension (30s), and a final extension of 72°C for 5min) and sent post-PCR processed samples (exoSAP-IT, usb catalog #78201) for sequencing (Genewiz, Inc., South Plainfield, NJ).

4.5.4 Affymetrix array analysis

Using an Affymetrix array containing 74,656 markers [111], we hybridized parasites with identical barcodes and parasites within the same collection but with different barcodes as well as technical replicates of control strains. We called SNPs using BRLMM-P from Affy Power Tools v1.10.2.Haploid genotypes were forced by designating all SNPs as "Y chromosome" and all individuals as "male". We counted the number of differing SNP genotypes for pairs of arrays, with pairings sorted into three categories: 1) technical replicates (same parasite sample hybridized to two arrays); 2) identical barcodes (distinct patient samples with identical barcodes); and, 3) unrelated parasites (distinct barcodes).

4.5.5 Human genotyping

We used a set of SNPs selected by The Broad Institute for human typing on their analysis platforms to distinguish patient samples from one another. From an original set of 23 assays, we selected 18 as robust under conditions with low template concentrations. We ran these pre-developed TaqMan-MGB probes (Life Technologies, Inc.) on an Applied Biosystems 7900HT qrt-PCR system (LifeTechnologies, Inc.) using the standard amplification and analysis protocols (see Supplementary Table 3 for SNP identity and human typing results).

In addition, we sent several samples for STR genotyping on an ABI 3130 Genetic Analyzer to detect the STR alleles amplified using the ABI AmpFlSTR Profiler Plus Kits (Life Technologies catalog # 4303326) at the Histocompatibility and Tissue Typing Laboratory, Brigham and Women's Hospital, Boston, MA. See Supplementary Table 4 for results of this genotyping.

4.5.6 Data analysis

We excluded from analysis those samples with missing data on more than four SNP positions. We determined that samples with more than one site showing both fluorescent signals in genotyping (indicating that more than one allele were present) were mixed infections with more than one genome present in the patient sample. For simplicity, the results we show in the paper are all based on samples with single genome. We also considered mixed infection in the analyses, and the results do not change qualitatively.

We calculated the standardized index of association (I_A^S) by the program LIAN, version 3.5 [48]. The number of re-samplings was set to be 10,000. We assumed there are two generations per year and estimated variance effective population size through temporal changes in allele frequencies by both the moment method [114] and likelihood approximation implemented in program CoNe [5]. We calculated the ratio of parasites persisting between years in each year through dividing the number of barcodes that are shared with other years by the total number of barcodes in a particular year.

5

Human cerebral malaria and *Plasmodium falciparum* genotypes in Malawi

Besides studying and tracking parasite populations within geographical regions, we explored the parasite populations within single patients. Using the molecular barcode described in Chapter 4 and [24], we applied the TaqMan SNP assays to parasite material extracted from multiple tissue types from autopsy patients diagnosed with cerebral malaria as well as a control set of patients who had died from other presumptive causes.

We found that the parasite types extracted from individual tissues matched those found in the peripheral blood from the same patient, indicating first that peripheral blood has utility for tracking infections in more complicated severe malaria cases. The statistically significant finding of single-genome infections in the cerebral malaria cases, as well as significant association with retinopathy-positive diagnosis, indicate that single and low-complexity *Plasmodium falciparum* populations dominate severe malaria infections.

Further, in temporal studies of this relatively high-transmission region, we saw no evidence of clonal barcode propagation, with no shared parasite types appearing within or between seasons.

5.1 Introduction

Patients suffering malaria can have uncomplicated malaria, where the disease is curable when treated promptly with appropriate anti-malaria treatments and with no long-term complications. However, a number of cases are severe malaria, especially in immunologically naïve adults and children under the age of 2. Severe malaria is generally diagnosed when the patient suffers acute and severe symptoms such as cardiovascular or respiratory distress and symptoms associated with massive lysis of red blood cells [122]. The WHO defines cerebral malaria, a special case of severe malaria, as a patient in an unresponsive and otherwise unexplained coma for at least one hour (for children) with accompanying parasitemia in the peripheral blood. This clinical definition, however, has been recently refined to include detectable presence of malarial retinopathy [11]. Approximately 20% of cerebral malaria victims die, and a number of the survivors suffer permanent disabilities from the neurological effects.

While progress has been made towards better diagnoses of severe and cerebral malaria infections, less is understood about the parasites that lead to these conditions. A classic symptom of cerebral malaria upon autopsy is the sequestration of non-sexual-stage parasites in the blood vessels of the brain; however, the mechanism for this phenomenon is still under investigation. This is in contrast to non-severe malaria cases, which generally do not feature organ sequestration [109].

Several studies have utilized merozoite surface protein (*msp*) typing of highly polymorphic antigenic regions of the parasite to conclude generally that as infections become more severe, the complexity of infection decreases [20, 21]; but these methodologies cannot identify parasite types.

In order to better understand the parasite population within patients with cerebral malaria compared to those with uncomplicated cases, we first tested the performance of the molecular barcode in the presence or absence of cerebral malaria in autopsy tissues where the diagnosis

was most accurate [75]. We next analyzed the molecular barcode in the peripheral blood of patients with clinically defined cerebral malaria, comparing results in patients with and without malarial retinopathy. We hypothesize that "true" cerebral malaria patients, meeting WHO clinical case definition and also malaria retinopathy positive, would demonstrate a strong association with single/low complexity infections in peripheral blood due to the homogenous total body parasite biomass and the ability of the molecular barcode to uniquely identify genotypes. We further hypothesize that non-cerebral malaria patients (clinical case definition and malaria retinopathy negative) patients would show various patterns (single or low-complexity or mixed) due to the presumed incidental nature of this population. We include important co-factors in this analysis including the effect of data of admission (such as time point in the malaria season) as we questioned what the effect of seasonal immunity would be on clinical infections. If this hypothesis is supported it will allow us to directly study with in depth sequencing the pathogenic strain in an individual patient, allowing for genome-wide association studies by identifying the critical parasite associated with disease episode. In addition, by analyzing the specific signature of the molecular barcode, we will be able to determine if a single genotype is responsible for cerebral malaria in this population.

5.2 Materials and methods

5.2.1 Definition of cerebral malaria

The WHO clinical definition of cerebral malaria (CM) includes the following: a Blantyre coma score ≤ 2, parasitemia by blood film, and no other evident cause of coma (e.g., meningitis, post-ictal state, hypoglycemia) [122].

The definitive diagnosis of cerebral malaria relies on post-mortem examination of the brain either by autopsy or supraorbital sampling. It is established that the clinical diagnosis of CM is strongly associated with the pathological finding of parasite sequestration (attributable to cytoadherence of parasites to endothelium) within the cerebral vasculature [72, 109]. Finding malarial retinopathy on ophthalmoscopy in a comatose pediatric patient supports a diagnosis of CM. [10, 11, 73, 116, 117]. The specific features of retinopathy that are indicative of a malarial cause of illness are vessel color changes and retinal whitening. White-centered hemorrhages are suggestive of malaria but are also seen in other conditions outside of the

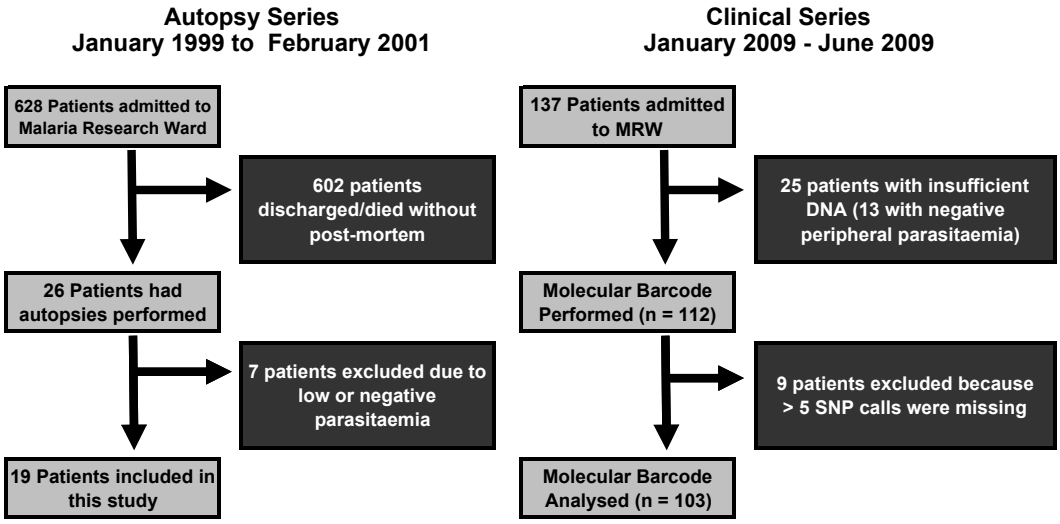


Figure 5.2.1: Flowchart of patients and samples collected for these studies.

setting of tropical pediatrics. Retinopathy is associated with intracerebral sequestration of parasites and with the severity and outcome of the disease [10].

5.2.2 Study design and patients

Two cohorts of patients were included in this study nested within the clinicopathological study in the Paediatric Research Ward (PRW, Queen Elizabeth Central Hospital, Blantyre, Malawi) (Figure 5.2.1). The autopsy series includes 19 autopsies (January 1999 to June 2001) in which we previously assessed genotypes (characterized by *msp*-1 and -2) of tissue-sequestered parasites [75]. We compared tissues collected at autopsy (six sites per patient including three brain sites, heart, lung, and colon) and peripheral blood at time of admission to the research ward (one per patient when available) for the definitive pathological diagnostic groups CM versus other causes of death. A total of 120 samples were available (19 x 6 tissues + 6 x 1 peripheral bloods).

The second cohort was comprised all patients (n = 137) admitted between January and June of 2009. Peripheral blood was collected on FTA cards (Whatman) at admission. All genotyping was performed blinded to patient information including parasitemia. We compared two groups (i.e., those with and without malaria retinopathy) specifically to under-

stand the peripheral blood genotypes associated with cerebral malaria as compared with parasites found in patients with incidental parasitemia. The use of retinopathy alone to separate CM from assumed incidental parasitemia will lead to some misclassification (specifically, false negatives); however, for the purposes of this study we assume complete delineation and interpret the results in this context. Patients meeting the clinical case definition without evidence of malaria retinopathy are a mixed group of diagnoses with essentially incidental parasitemia and no to limited histological evidence of sequestered parasite populations in the brain or other organs. In this clinical study, we did not readily have access to deep tissue sites of sequestered populations which may have been informative (such as the gastrointestinal tract which, in our autopsy data, suggests a strong correlation to brain sequestration). Our autopsy observations do no support the skin as a valid correlation with brain it is also culturally unacceptable in our patient population. Thus, in this analysis, we rely solely on retinopathy to confirm cerebral malaria. The research ethics committees at Michigan State University, the University of Liverpool, the University of Malawi College of Medicine, and the Brigham & Women's Hospital have approved all or appropriate portions of this study.

5.2.3 The Malaria Research Ward

The Malaria Research Ward (MRW) admits children, with informed parental consent, to a program of clinical care and detailed observational studies. Patients are admitted who fulfil clinical criteria for a variety of malarial and non-malarial diagnoses. Diagnostic criteria, clinical management, laboratory investigations and treatment protocols have been previously described [109]. Final clinical diagnoses were derived from data collected throughout the hospital stay. The presence or absence of malarial retinopathy, defined as the presence in one or both eyes of vessel color changes (orange vessels or vessel whitening), retinal whitening, and/or hemorrhages was assessed after admission using direct and indirect ophthalmoscopy [10, 11]. In the event of death, a Malawian clinician or nurse met with key family members to request their consent for an autopsy. If permission was granted, the post-mortem was performed as quickly as possible in the mortuary at the Queen Elizabeth Central Hospital (typically, less than 12 hours).

5.2.4 Autopsy procedures

Gross examination, documentation, and histological assessment of the brains and other organs were performed, and a final anatomic diagnosis was determined as previously described [5]. Briefly, patients meeting the WHO clinical case definition of CM during life who were found to have sequestration of parasites in their brain were classified as CM; patients with these features plus a hematocrit less than 15% were classified as CM + severe malaria anemia (SMA); patients with a hematocrit of less than 15% and no other pathology were classified as SMA; and all other patients were classified by the anatomic cause of death (e.g., pneumonia or other).

5.2.5 DNA extraction

For the 20 autopsy patients included in this study, 200 uL of peripheral blood and 0.5 g of frozen tissue from six organ sites (frontal lobe, midbrain, cerebellum, lung, heart, colon) were used for DNA extraction by a previously described phenol:chloroform extraction protocol [75]. From the 137 admitted patients in the second part of the study, FTA peripheral blood samples were used for DNA extraction using QIAmp DNA Blood Mini Kit (Qiagen Catalog # 51106) using three-6 mm punches.

5.2.6 DNA quantification

We optimized the previously described parasite DNA quantification assay [24] for both a 96-well and a 384-well plate RT-PCR system. For the quantification of peripheral blood samples of admitted patients (MLW Research Laboratories (MLWRL), Blantyre, Malawi), where a 96-well plate system was available, a master mixture was prepared using 5.0 μ Master Mix (Applied Biosystems Catalog #4364343) and 0.5 μl of 20x PF07-0076 pre-mixed quantification assay. Experimental internal control for standard curve (3D7 from culture in serial dilution, verified by Nanodrop quantification) and samples were loaded into 96-well PCR plates (total volume of DNA and water was 5.0 μl in a 10 μl reaction) followed by addition of the master mixture. PCR conditions and analysis were as described previously [24].

5.2.7 Genotyping

The 120 organ samples and 6 peripheral blood samples from the 20 autopsy patients underwent molecular barcoding using the 24-SNP assay in a 384-well format as previously described [24] DNA extracted from the 112 peripheral blood samples from admitted patients underwent molecular barcoding using a 24-SNP assay in a 96-well format performed in the field (MLWRL, Blantyre, Malawi) as follows: template DNA and water in a total volume of 5.0 μl was added to a 5.0 μl mix made up of 0.250 μl 40x SNP assay and 5.0 μl Master Mix (AB Catalog # 4364343) in a 96-well optical PCR plate and mixed, for a total reaction volume of 10 μl. The PCR amplification conditions and analytical approach were not changed [24]. For all barcodes, raw data and allelic calls were made blinded to all clinical data and independently by at least two observers and discrepancies were resolved by consensus.

5.2.8 Molecular barcode interpretation

For each SNP call, the four possible results include: allele 1 is present; allele 2 is present; both alleles are present (heterozygous); or the assay fails. Because the parasites that are being sampled are in the intra-erythrocytic stage of their life cycle and are, therefore, haploid, identifying both alleles present in a single sample means, by definition, that there must be at least two genomes. In the development of the 24-SNP molecular barcode assay, ratio experiments using known mixtures of two different single clone parasites were performed revealing the minimum ratio of individual assays (5:1 17%; 10:1 65%; 20:1 21%) which supports, in vitro, that a single barcode signature indicates at least 90% of the DNA content was from a single genome[117]. When all 24 alleles are single calls (i.e., no heterozygous calls), this suggests a single genotype is present at the 90% or greater level. Autopsy samples were analyzed by diagnosis and classified based on the previous *msp-1* and *-2* data and the molecular barcodes from the autopsy data (Table S1). Peripheral blood samples from the second group were classified based solely on molecular barcode using assumptions from autopsy data. For this analysis, 0, 1, or 2 heterozygous calls were considered single/low complexity infections while 3 or more heterozygous calls were considered mixed infections. In our previous work, failed reactions were always due to either absent genomic DNA (deletions) or insufficient DNA quantity in the sample.

5.2.9 Statistical analysis

Comparisons of baseline characteristics between patients at autopsy and clinical patients with and without retinopathy were based on t-tests or Wilcoxon tests when studying numerical characteristics. We explored the effect of date of admission on heterozygosity using an over-dispersed Poisson regression, since we expected that multiplicity of infection would decrease as the malaria transmission season progressed owing to increasing immunity level and decreasing transmission pressure. Logistic regression, unadjusted and adjusted for potential confounders, was performed to study the effect of presence of a single/low-complexity infection (as opposed to a complex population) on malaria retinopathy. Results of those models were compared to results of models including number of heterozygous calls, i.e., studying the impact of one unit increase in heterozygous calls on the likelihood of malaria retinopathy. Candidate confounders included patient s age, parasite density, hematocrit, and the date on which the sample was collected. These were retained in models when the P-value of their likelihood-likelihood ratio test was < 0.05 or when they appreciably impacted the coefficient of the main variable of interest less complex vs. more complex population. When modeling numeric variables, linearity of associations was assessed using splines in generalized additive models. No variable had a statistically significant departure of linearity. Adding non-linear terms for the selected candidate confounders did not appreciably change the odds ratio of the main predictor of interest. All statistics and graphs were analyzed and/or created with STATA v9.0 (College Station, Texas), R (www.r-project.org), and Graphpad Prism v4.03 (La Jolla, California). Statistical tests were considered significant to an α-level of 0.05.

5.3 Results

5.3.1 Retrospective autopsy study: limited parasite complexity in brain and tissue of cerebral malaria patients

Autopsy data from 19 patients were examined including five with CM, five with CM+SMA, one SMA, four pneumonias, and four other non-malarial diagnoses. We analyzed autopsy material and determined the genotypes of parasites sequestered in tissues and circulating in the peripheral blood (Figure 5.2.1 & Table 5.3.1). We found that molecular barcodes across tissues within an individual were nearly identical (parasites found in the heart, lungs, colon,

Table 5.3.1: Comparison of baseline characteristics. (A) 19 autopsy patients presented in the first part of the study and (B) the admitted patients presenting with or without features of malaria retinopathy in the second part of the study

Characteristic	**Cerebral Malaria (+/− SMA)**	**Non-CM (including SMA)**	**P***
	(N = 10)	**(N = 9)**	
	Median (IQR)		
Age [months]	24 (18-70)	25.5 (17-39)	0.64
Time [weeks]†	13.6 (9.4 - 16.4)	10 (9-14.3)	0.49
Haematocrit [%]	18 (14-31)	29.5 (22-40)	0.15
Parasitaemia [p/ul]	275,200 (11,399-572,880)	98,518 (4,716-286,850)	0.25
Characteristic	Retinopathy Positive (N = 69)	Retinopathy Negative (N = 34)	P*
	Median (IQR)		
Age [months]	38 (25-56)	51 (31-67)	0.19
Time [weeks]†	7.4 (4.3, 12.9)	10.4 (4.1, 17.4)	0.44
Haematocrit [%]	22 (18, 27)	30 (25, 34)	0.0003
Parasitaemia [p/ul]	101,176 (31,590-366,070)	70,230 (42,871-211,680)	0.71

and, when available, peripheral blood, were the same by the assay as those found in the brain Figure 5.3.1 & Supplementary Figure 7) in patients with retinopathy-positive CM (with or without severe anaemia). It is important to note that this included some mixed infections (many heterozygous calls) that were still conserved across tissues suggested the total body population was the same mixture, an observation which was seen primarily in the CM+SMA patients. More complex infections (variation from consensus across tissues) were found more commonly in the patients without evidence of retinopathy who had a range of other demonstrable causes of death. In addition, the peripheral blood reflected the extent of parasite complexity found in tissues (few genotypes in blood equals few genotypes in tissue and vice versa) (Supplementary Figure 7).

5.3.2 Prospective clinical study: limited parasite complexity in peripheral blood of patients with cerebral malaria

A total of 137 patients had peripheral blood collected on FTA cards and 25 patients were excluded because they had either insufficient DNA by quantification or they had negative peripheral parasitemia. For 112 patient samples that were barcoded, a total of 2736 individ-

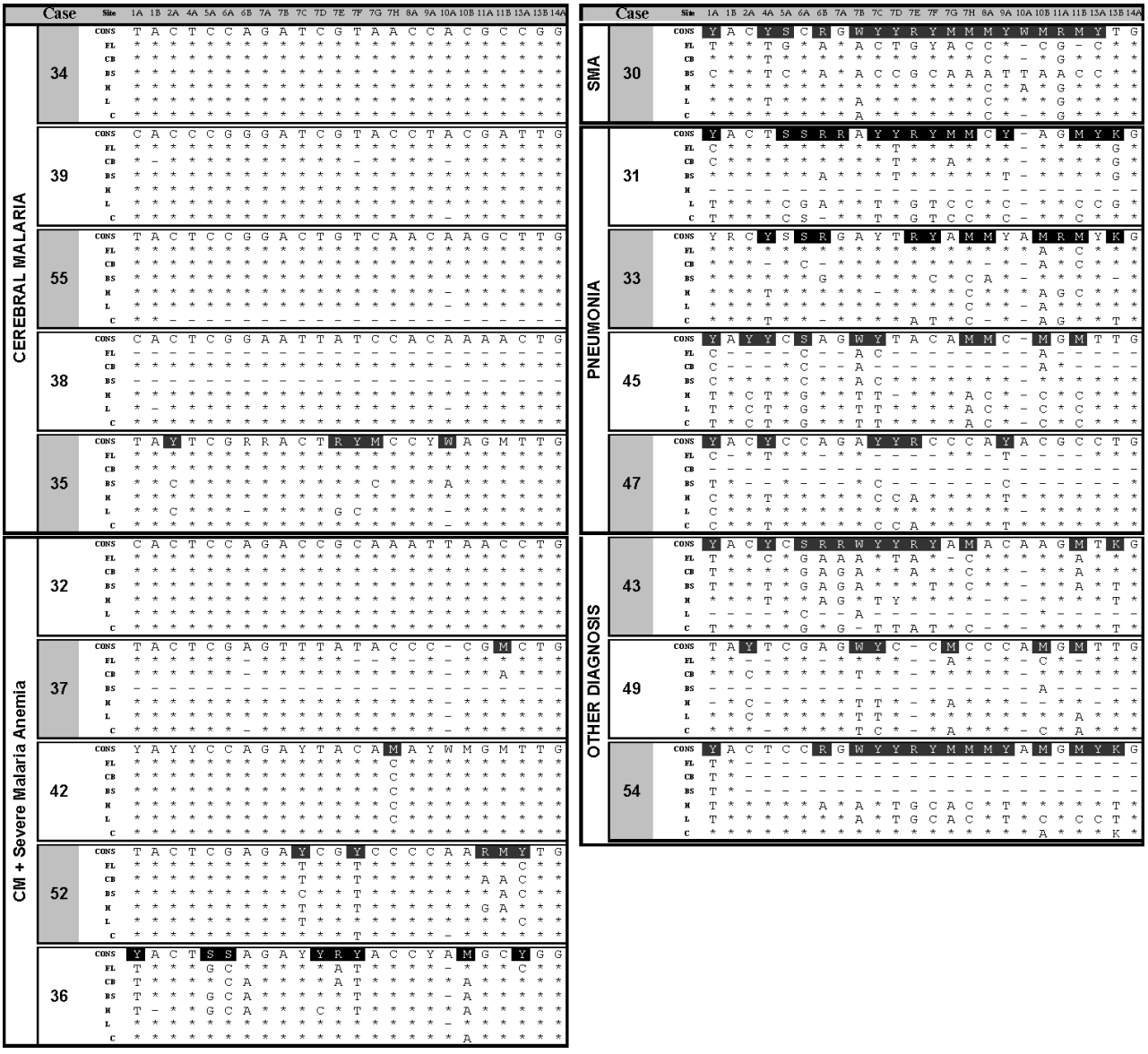


Figure 5.3.1: Molecular barcodes for 18 autopsy patients. Consensus barcode in the first row and individual barcodes for frontal lobe (FL), cerebellum (CB), brainstem (BS), heart (H), lung (L), and colon (C). Among the eight non-CM patients, the diagnoses were SMA (1), pneumonia (3), sepsis (1), giant cell myocarditis (1), ruptured arteriovenous malformation (1), and traumatic skull fracture (1). Where the individual barcode matches consensus, an * is used. Where the individual barcode differs from consensus, the reported allele is given (A, G, C, or T). When the consensus barcode call is not conserved across all tissues, the consensus call is highlighted in black with white text. For samples which were either not available or the reaction failed, a dash (-) is used to denote missing. For a heterozygous consensus call, the IUPAC code (Supplemental Table 7) is used and BOTH bases denoted by the code are present.

Table 5.3.2: Comparison of the major alleles encountered in the Malawi consecutive patient data set. One season (n = 112) compared with the same allele for the set of global parasites previously published (n = 114).

Assay	Malawi	Original	Difference	z-score	*p*-value
1	0.6446	0.8690	0.2244	-7.1029	<0.05
2	0.9099	0.8475	0.0625	1.8546	NS
3	0.6591	0.6885	0.0294	-0.6786	NS
4	0.8182	0.6446	0.1736	3.8716	<0.05
5	0.7653	0.7698	0.0045	-0.1150	NS
6	0.6269	0.6154	0.0115	0.2520	NS
7	0.6270	0.4640	0.1630	3.4894	<0.05
8	0.8130	0.8034	0.0096	0.2576	NS
9	0.7481	0.5873	0.1608	3.4871	<0.05
10	0.6391	0.6111	0.0280	0.6130	NS
11	0.6529	0.3220	0.3309	7.5603	<0.05
12	0.5620	0.4274	0.1346	2.9058	<0.05
13	0.6504	0.5289	0.1215	2.5985	<0.05
14	0.7241	0.6239	0.1002	2.2087	<0.05
15	0.5489	0.5124	0.0365	0.7791	NS
16	0.7652	0.4872	0.2780	5.9392	<0.05
17	0.5556	0.6032	0.0476	-1.0392	NS
18	0.6330	0.5124	0.1206	2.5768	<0.05
19	0.5909	0.6250	0.0341	-0.7519	NS
20	0.6692	0.6230	0.0462	1.0183	NS
21	0.5231	0.4318	0.0913	1.9671	<0.05
22	0.5766	0.3953	0.1813	3.9591	<0.05
23	0.6866	0.6518	0.0348	0.7795	NS
24	0.9231	0.5897	0.3333	7.2356	<0.05

Thirteen assays are significantly different from the population observed in the original study with the differences ranging from 9.1 to 41.0%. Assay 24 is non-informative in the Malawi population

ual SNP assays were performed with 207 reaction failures (7.6%) and 466 heterozygous calls (17%); the baseline allele frequencies were different between the Malawi population and the global population previously studied with one assay (SNP 14A) being non-informative in the Malawi samples (Table 5.3.2 and Supplementary Table 7)[24]. The molecular barcodes of 103 patients were analyzed, 69 with and 34 without malaria retinopathy; 9 samples were excluded due to greater than 5 failed reactions out of 24. Fifty-two samples contained single or less complex infections (0, 1, or 2 heterozygous calls) while 51 samples had more complex

populations ($\geq$ three heterozygous calls, maximum number observed in a sample = 19). When comparing retinopathy positive to retinopathy negative patients, 77% of retinopathy positive patients had singleor low-complexity infections including 28 samples with 0 heterozygous calls suggesting a single genotype (Figure 5.3.2). The unadjusted OR for having positive malarial retinopathy in individuals with a low complex infection was 3.70. We performed a sensitivity analysis using only molecular barcodes with 0 failures, 1 failure, etc. up to 5 failures within an individual barcode and found the same unadjusted odds ratio (3.7 to 4.2) which was significant (P-value < 0.01) except for the case of 0 failures due to sample size (P-value = 0.071). Patients with and without malaria retinopathy were clinically similar except for hematocrit (Table 5.3.1). When adjusted for hematocrit, parasite density (confounding variable), and the effect of time, the OR for a single/less complex (vs. more complex) infection having retinopathy was 4.8 (Table 5.3.3).

Although we saw a preponderance of single or less-complex parasite populations in cerebral malaria patients, we found no unique genotype or group of genotypes associated with cerebral malaria, nor did we find any instances of repeated barcodes between patients or between seasons. In addition, we found that the complexity of infections decreases over the six months during which data were collected (i.e., from start to end of the malaria season). The number of heterozygous calls per patient decreased over time and, compared to retinopathy negative patients collected concurrently, retinopathy positive patients had consistently and significantly fewer heterozygous calls per patient (p-value = 0.00007, Figure 5.3.3).

5.4 Conclusions and discussion

This study focused exclusively on children who meet the clinical case definition of cerebral malaria, all of whom had *P. falciparum* parasitaemia. Patients were separated solely on the presence (retinopathy-positive CM) or absence (retinopathy-negative CM) of malaria retinopathy to ask what the differences are in genetic components of infection between these two phenotypes using the molecular barcode technique. Genotyping by *msp*-1 and -2 has been used to evaluate these patients previously [75] and the current study supports and extends those observations to show a dominant genotype by barcoding.

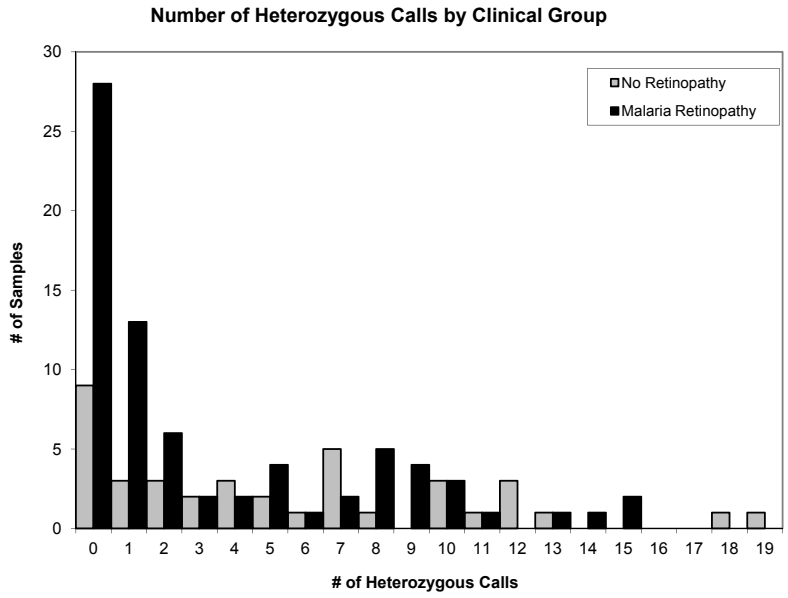


Figure 5.3.2: Number of heterozygous calls in molecular barcodes. Molecular barcodes were grouped by number of heterozygous calls (range = 0-19) after excluding barcodes missing in more than five assays and were then categorized by malaria retinopathy (presence or absence). A one unit decrease in heterozygous calls is associated with increased risk of presenting with malaria retinopathy: adjusted OR 1.11; 95% CI = 1.01, 1.22; $P = 0.03$.

Table 5.3.3: Association between malaria retinopathy and infections caused by a single/less complex genotypes (vs three or more).

	Single/Less Complex Genotype	**Multiple Genotypes**	
	(< 3 Het. Calls)	**(≥3 Het. Calls)**	
	(N = 52)	**(N = 51)**	
Retinopathy	42 (81%)	27 (53%)	
No retinopathy	10	24	
	Logistic Regression		
	(N = 67 with and 32 without retinopathy)		
Unadjusted Model†	**OR**	**(95% CI)**	**P***
Single/Less Complex (vs. 3 Hets)	3.7	(1.51-9.10)	0.003
Adjusted Model†			
Single/Less Complex (vs. ≥ 3 Hets)	4.82	(1.59-14.30)	0.003
Hematocrit	0.84	(0.79, 0.96)	< 0.0001
Parasites/μl (natural logarithm)	1.36	(0.996, 1.85)	0.046
Time (weeks from January until date of sample collection)	0.87	(0.77-0.92)	0.001

Frequencies and odds ratios (OR) obtained in logistic regression models. Four patients were removed from regression modeling due to missing data
Het. = Heterozygous; *OR* = Odd Ratio; *CI* = Confidence Interval
*P-values were estimated using likelihood-ratio tests.
†c-statistic (corresponding to the area under the ROC curve of the fitted model) of the unadjusted model was 0.66 and of the adjusted model was 0.82.

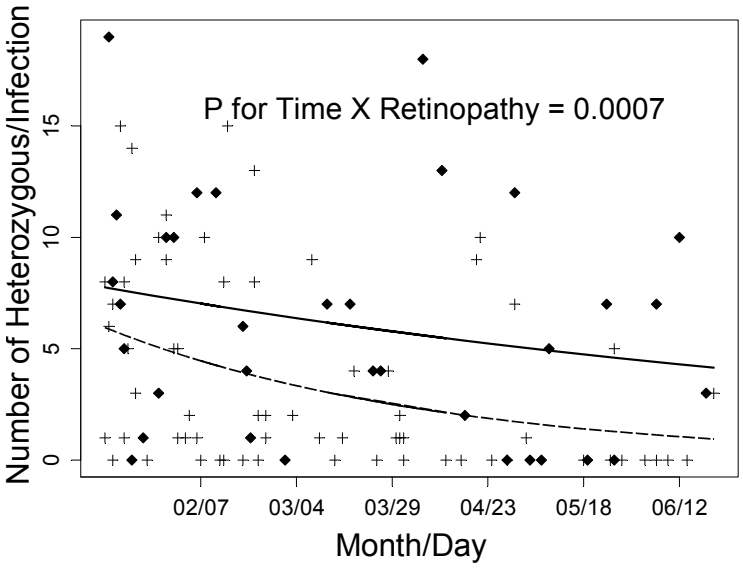


Figure 5.3.3: Changes in the number of heterozygous calls per barcode in individual patients (Y-axis) as days of sample collection progressed (X-axis) January to June 2009. The crosses and dashed line (obtained from an over-dispersed Poisson model) represent the malaria retinopathy-positive patients with a decrease of approximately 5.0 heterozygous calls over six months. The diamonds and solid line (obtained in an over-dispersed Poisson model) represent the malaria retinopathy-negative patients with a less marked decrease of approximately 3.6 heterozygous calls over six months. The trend suggests that over the course of the data collection period (i.e. the malaria season in Malawi), the number of mixed infections decreases leading to more homogenous infections. Trends over time of patients with and without malaria retinopathy were statistically significantly different (P-value = 0.0007).

Previous studies have shown that malaria retinopathy is strongly associated with pathological confirmation of cerebral parasite sequestration, while parasitaemic comatose patients without malaria retinopathy represent a diverse group of non-malarial diseases [109, 117]. This current data support the concept of limited parasite diversity in cerebral malaria patients and are consistent with mathematical models and immunological evidence from a broad range of studies [15, 22, 37, 41–45, 59, 61, 88, 93]. These results validate the previous *msp*- 1 and *msp*-2 data; importantly, they demonstrate that the molecular barcodes from peripheral parasitaemia mirror those generated from sequestered parasites in various tissues and this genotype is predominant in CM [75].

The number of heterozygous calls within individual patients with retinopathy-positive CM decreased over the malaria season, suggesting that complexity of infection is a product of time. This could be explained by a decreasing rate of parasite inoculations, or an increasingly broad acquisition of immunity in the population as the season progresses. Clinically evident infections may, as immunity accumulates, tend to occur only in patients inoculated with new genotypes for which no immunity has developed.

Parasites from peripheral blood and parasites found in tissues at autopsy had the same conserved barcode identification within an individual patient in CM patients. Less complex infections in the peripheral blood were more common in patients with stringently characterized cerebral malaria (standard clinical case definition plus retinal examination) and include apparently single genotypes fairly frequently (35%). However, there was not a single global conserved barcode genotype associated with cerebral malaria, and no instances of barcode genotype appearing within any other patients. The tissues of patients without cerebral malaria showed different signals in different tis- sues, and this heterogeneity is reflected in the peripheral blood samples. The corresponding histological tissue slides from these patients show very few sequestered parasites suggesting that the signal from individual tissues came from either small collections of sequestered parasites or parasite DNA within phagocytic macrophages. Cerebral malaria patients had less complex infections in tissue (average of 1.6 heterozygous calls per sample = two to three genotypes per sample), a finding consistent with the earlier study in which the complexity was characterized using *msp*-1 and *msp*-2 genotyping (average of 1.9 genotypes per sample) and with the literature (Supplementary Figure 6) [75].

What remains unknown is the full range of diversity within a single cerebral malaria

patient with such an enormous sequestered burden. The autopsy data presented here are from patients who died from the illness; thus, single/low complexity infections may be more likely to result in death. The observations made in this clinical cohort suggest that a wider range of diversity is possible within the individual with or without retinopathy, although in these clinical patients those with retinopathy had significantly less complex infections than those without.

6

Concluding remarks and future directions

In a new era of hope for malaria eradication, we rely on better information to design effective strategies based on the combined work of clinical and laboratory researchers as well as front-line clinicians, health care workers and governmental bodies. Bolstering these efforts has been a data renaissance: increased access to sequencing and other genetic information that has the potential to answer a plethora of biological questions. Of particular interest is the population structure of the malaria parasites and how it changes through natural and selective pressures.

Given the population genetic fundamentals reviewed and applied to *Plasmodium* in Chapter 2, this work used distinct sets of markers to track population genetics of *P. falciparum*. First, using markers expected to be under selection, we developed and applied assays to detect SNPs associated with reduced drug sensitivity. The data discussed in Chapter 3 revealed the changing ratios of mutations within different populations, detected the emergence of a novel mutation, and pinpointed the emergence or importation into the Senegalese parasite population of a SNP reported elsewhere.

Then, in Chapter 4, we used neutral markers to study temporal trends in *P. falciparum* infections from patients being evaluated at the SLAP clinic in Thiès, Senegal. From 2006 to 2011 we identified a trend of increasingly less complex infections, with an increased proportion of single infections, suggesting clonal propagation of the parasite due to reduced outcrossing in the mosquito midgut. Furthermore, we discovered evidence of epidemic expansion of certain parasite types, with 25-30% of the single infections in 2008 and 2011 containing the same parasite molecular barcode. Coincident with this finding was the observation that the parasite types persisted across years and dry seasons, sometimes emerging intact over several years.

Chapter 5 discusses a more narrow focus of the molecular barcode to study the population of parasites within children suffering from cerebral malaria. This technique was able to show that the complexity of infection in patients with severe and cerebral malaria was generally less complex than patients without clinical diagnosis of cerebral malaria. Furthermore, the parasites found in peripheral blood matched those sequestered in organs, making it an acceptable proxy for parasites in those organs. No single or group of parasites were found in prevalence in the CM patients or in the population as a whole.

These Chapters describe the tracking of *Plasmodium falciparum* parasite populations using a variety of markers. What makes these approaches particularly powerful is that they offer a real translation from laboratory research to field surveys. While sequencing of parasite genomes directly from patient samples is becoming more technically feasible and approachable for more budgets [70], the costs are still too great and the analysis of the material requires more skill than most settings can afford. The development and dissemination of facile tools for surveillance of population changes empowers real-time observation in more settings to answer outstanding questions; for instance, few studies track parasite populations over time outside of the spread of drug resistance [13, 28, 39, 74, 87, 100]. Those few groups, including ours, that have described temporal studies observe real and significant changes that must surely effect control program decisions; however, the limited number of studies emphasizes the dearth of information about parasite populations and their potential responses to control efforts.

In particular, they highlight the necessity to understand the population dynamics as transmission is decreased and the parasite population changes, which will lead to better assessment of control and pre-elimination status and also allow better prediction of the likelihood

of program success. In particular, as the number of cases decreases, appropriate population surveillance is necessary to detect asymptomatic patients and travelers as potential reservoirs of the parasite.

Easily-implemented methods for this surveillance are necessary, as well as improved methods for more fine-scale tracking of populations. The high-resolution melting assays described are just a first step; with additional information from genome-wide association studies (GWAS) to highlight SNPs associated with reduced drug sensitivity, assays can be quickly designed and implemented for tracking of these emerging resistances.

The molecular barcode currently gives a sentinel view of the population and detects large-scale changes associated with significant reduction in effective population size; however, with additional markers throughout the genome and centered in regions reported to be hotspots for recombination [46], the utility of the barcode could be expanded to detect earlier hallmark changes in population such as reductions in relative complexity of infection.

The work described in Chapter 5 particularly highlights the complexities of surveillance and discovery in sites with higher transmission and concomitantly higher COI. This challenge is one for which the current barcode is not yet optimized. While the molecular barcode has shown itself to be valuable in regions with low transmission, additional work is required to better understand areas like Malawi to better answer population genetic questions where complexity of infection is high. With additional information about the nature of these mixed infections, we can evaluate the proposed predictor of multi-clonality presented in the Chapter. Sequencing is one way to address these questions most directly, but single cell capture and amplification will also provide information about parasites on an individual level without potential confounders of sequencing error and biased reads still common with existing sequencing technologies, especially when applied to mixed infections [70].

Further work is needed to more directly link these genomic observations to trends and reports in transmission as well as the clinical instances and description of disease. Where program re-orientation is required at specific epidemiological milestones, we need to better understand if those milestones correspond to the population changes that we have described. In particular, as transmission decreases, the dynamics of vector mating are particularly important to better understand trends with mixed gametes, reassortment, and recombination contributing to the parasite population structure.

As the most virulent and deadly of the malaria species, *Plasmodium falciparum* remains the priority for most control programs and international efforts; however, the methodologies and research requirements developed for *P. falciparum* will become essential when addressing the next-highest cause of morbidity, *P. vivax*. These studies pave the way for additional work in these other organisms.

References

[1] Fisher R A. *The genetical theory of natural selection.* Clarendon Press, Oxford, 1930.

[2] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, September 2009.

[3] Alfred Amambua-Ngwa, Daniel J Park, Sarah K Volkman, Kayla G Barnes, Amy K Bei, Amanda K Lukens, Papa Sene, Daria Van Tyne, Daouda Ndiaye, Dyann F Wirth, David J Conway, Daniel E Neafsey, and Stephen F Schaffner. SNP genotyping identifies new signatures of selection in a deep sample of West African Plasmodium falciparum malaria parasites. *Molecular Biology and Evolution*, 29(11):3249–3253, November 2012.

[4] William Amos. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proceedings of the Royal Society B: Biological Sciences*, 277(1686):1443–1449, May 2010.

[5] Eric C Anderson. An efficient Monte Carlo method for estimating Ne from temporally spaced samples using a coalescent-based likelihood. *Genetics*, 170(2):955–967, June 2005.

[6] T J Anderson, B Haubold, J T Williams, J G Estrada-Franco, L Richardson, R Mollinedo, M Bockarie, J Mokili, S Mharakurwa, N French, J Whitworth, I D Velez, A H Brockman, F Nosten, M U Ferreira, and K P Day. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. *Molecular Biology and Evolution*, 17(10):1467–1482, October 2000.

[7] Valérie Andriantsoanirina, Vincent Lascombes, Arsène Ratsimbasoa, Christiane Bouchier, Jonathan Hoffman, Magali Tichit, Leon-Paul Rabarijaona, Rémy Durand, and Didier Ménard. Rapid detection of point mutations in Plasmodium falciparum genes associated with antimalarial drugs resistance by using High-Resolution Melting analysis. *Journal of Microbiological Methods*, 78(2):165–170, August 2009.

[8] Mary Lynn Baniecki, Dyann F Wirth, and Jon Clardy. High-throughput Plasmodium falciparum growth assay for malaria drug discovery. *Antimicrobial Agents and Chemotherapy*, 51(2):716–723, February 2007.

[9] L K Basco, O Ramiliarisoa, and J Le Bras. In vitro activity of pyrimethamine, cycloguanil, and other antimalarial drugs against African isolates and clones of Plasmodium falciparum. *The American Journal of Tropical Medicine and Hygiene*, 50(2): 193–199, February 1994.

[10] Nicholas A NA Beare, Caroline C Southern, Chipo C Chalira, Terrie E TE Taylor, Malcolm E ME Molyneux, and Simon P SP Harding. Prognostic significance and course of retinopathy in children with severe malaria. *Archives of Ophthalmology*, 122 (8):1141–1147, August 2004.

[11] Nicholas A V NA Beare, Terrie E TE Taylor, Simon P SP Harding, Susan S Lewallen, and Malcolm E ME Molyneux. Malarial retinopathy: a newly established diagnostic sign in severe malaria. *The American Journal of Tropical Medicine and Hygiene*, 75 (5):790–797, November 2006.

[12] K Beshir, C J Sutherland, I Merinopoulos, N Durrani, T Leslie, M Rowland, and R L Hallett. Amodiaquine Resistance in Plasmodium falciparum Malaria in Afghanistan Is Associated with the pfcrt SVMNT Allele at Codons 72 to 76. *Antimicrobial Agents and Chemotherapy*, 54(9):3714–3716, August 2010.

[13] OraLee H Branch, Patrick L Sutton, Carmen Barnes, Juan Carlos Castro, Julie Hussin, Philip Awadalla, and Gisely Hijar. Plasmodium falciparum genetic diversity maintained and amplified over 5 years of a low transmission endemic in the Peruvian Amazon. *Molecular Biology and Evolution*, 28(7):1973–1986, July 2011.

[14] D R Brooks, P Wang, M Read, W M Watkins, P F Sims, and J E Hyde. Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, Plasmodium falciparum, with differing resistance to sulfadoxine. *European Journal of Biochemistry / FEBS*, 224(2):397–405, September 1994.

[15] P C Bull, B S Lowe, M Kortok, C S Molyneux, C I Newbold, and K Marsh. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Medicine*, 4(3):358–360, March 1998.

[16] Centers for Disease Control and Prevention. The History of Malaria, an Ancient Disease. URL http://www.cdc.gov/malaria/about/history/.

[17] Hsiao-Han Chang, Daniel J Park, Kevin J Galinsky, Stephen F Schaffner, Daouda Ndiaye, Omar Ndir, Souleymane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Molecular Biology and Evolution*, 29(11):3427–3439, November 2012.

[18] Ian H Cheeseman, Natalia Gomez-Escobar, Celine K Carret, Alasdair Ivens, Lindsay B Stewart, Kevin K A Tetteh, and David J Conway. Gene copy number variation throughout the Plasmodium falciparum genome. *BMC Genomics*, 10:353–353, August 2009.

[19] Sandrine Cojean, Véronique Hubert, Jacques Le Bras, and Rémy Durand. Resistance to dihydroartemisinin. *Emerging Infectious Diseases*, 12(11):1798–1799, November 2006.

[20] J M Colborn, O A Koita, M W Bagayoko, and D J Krogstad. Emergence of new genotypes and increases in dominant genotype copy number are associated with development of symptomatic malaria in the village of Missira, Mali. . *The American Journal of Tropical Medicine and Hygiene*, April 2006.

[21] James M JM Colborn, Ousmane A OA Koita, Ousmane O Cissé, Mamadou W MW Bagayoko, Edward J EJ Guthrie, and Donald J DJ Krogstad. Identifying and quantifying genotypes in polyclonal infections due to single species. *Emerging Infectious Diseases*, 12(3):475–482, March 2006.

[22] A A Craig and A A Scherf. Molecules on the surface of the Plasmodium falciparum infected erythrocyte and their role in malaria pathogenesis and immune evasion. *Molecular and biochemical parasitology*, 115(2):129–143, July 2001.

[23] R E Cruz, S E Shokoples, D P Manage, and S K Yanow. High-Throughput Genotyping of Single Nucleotide Polymorphisms in the Plasmodium falciparum dhfr Gene by Asymmetric PCR and Melt-Curve Analysis. *Journal of Clinical Microbiology*, 48(9): 3081–3087, August 2010.

[24] Rachel Daniels, Sarah K Volkman, Danny A Milner, Nira Mahesh, Daniel E Neafsey, Daniel J Park, David Rosen, Elaine Angelino, Pardis C Sabeti, Dyann F Wirth, and Roger C Wiegand. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malaria Journal*, 7:223, October 2008.

[25] Sabelo V Dlamini, Khalid Beshir, and Colin J Sutherland. Markers of anti-malarial drug resistance in Plasmodium falciparum isolates from Swaziland: identification of pfmdr1-86F in natural parasite isolates. *Malaria Journal*, 9(1):68, March 2010.

[26] Zachary Dwight, Robert Palais, and Carl T Wittwer. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics (Oxford, England)*, 27(7):1019–1020, April 2011.

[27] Richard T Eastman, Neekesh V Dharia, Elizabeth A Winzeler, and David A Fidock. Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured Plasmodium falciparum parasites. *Antimicrobial Agents and Chemotherapy*, 55(8):3908–3916, August 2011.

[28] Diego F Echeverry, Shalini Nair, Lyda Osorio, Sanjay Menon, Claribel Murillo, and Tim Jc Anderson. Long term persistence of clonal malaria parasite Plasmodium falciparum lineages in the Colombian Pacific region. *BMC Genetics*, 14:2, 2013.

[29] J C Fay and C I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155 (3):1405–1413, July 2000.

[30] David A Fidock, Takashi Nomura, Angela K Talley, Roland A Cooper, Sergey M Dzekunov, Michael T Ferdig, Lyann M B Ursos, Amar bir Singh Sidhu, Bronwen Naudé, Kirk W Deitsch, Xin-Zhuan Su, John C Wootton, Paul D Roepe, and Thomas E Wellems. Mutations in the P. falciparum Digestive Vacuole Transmembrane Protein PfCRT and Evidence for Their Role in Chloroquine Resistance. *Molecular Cell*, 6(4): 861–871, October 2000.

[31] S J Foote, J K Thompson, A F Cowman, and D J Kemp. Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. falciparum. *Cell*, 57(6):921–930, June 1989.

[32] Y X Fu and W H Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3): 693–709, March 1993.

[33] Daniel J Gaffney and Peter D Keightley. The scale of mutational variation in the murid genome. *Genes & Development*, 15(8):1086–1094, August 2005.

[34] J L Gallup and J D Sachs. The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, 64(1-2 Suppl):85–96, January 2001.

[35] Linda Gan and Jin Loh. Rapid identification of chloroquine and atovaquone drug resistance in Plasmodium falciparum using high-resolution melt polymerase chain reaction. *Malaria Journal*, 9(1):134, May 2010.

[36] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, October 2002.

[37] H A Giha, T Staalsoe, D Dodoo, I M Elhassan, C Roper, G M Satti, D E Arnot, T G Theander, and L Hviid. Nine-year longitudinal study of antibodies to variant

antigens on the surface of Plasmodium falciparum-infected erythrocytes. *Infection and Immunity*, 67(8):4092–4098, August 1999.

[38] J P Gil, F Nogueira, J Strömberg-Nörklit, J Lindberg, M Carrolo, C Casimiro, D Lopes, A P Arez, P V Cravo, and V E Rosário. Detection of atovaquone and Malarone resistance conferring mutations in Plasmodium falciparum cytochrome b gene (cytb). *Molecular and Cellular Probes*, 17(2-3):85–89, March 2003.

[39] Sean M Griffing, Tonya Mixson-Hayden, Sankar Sridaran, Md Tauqeer Alam, Andrea M McCollum, César Cabezas, Wilmer Marquiño Quezada, John W Barnwell, Alexandre Macedo De Oliveira, Carmen Lucas, Nancy Arrospide, Ananias A Escalante, David J Bacon, and Venkatachalam Udhayakumar. South American Plasmodium falciparum after the malaria eradication era: clonal population expansion and survival of the fittest hybrids. *PLoS ONE*, 6(9):e23486, 2011.

[40] Sharon R Grossman, Ilya Shlyakhter, Ilya Shylakhter, Elinor K Karlsson, Elizabeth H Byrne, Shannon Morales, Gabriel Frieden, Elizabeth Hostetter, Elaine Angelino, Manuel Garber, Or Zuk, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–886, February 2010.

[41] S Gupta and R M Anderson. Population structure of pathogens: the role of immune selection. *Parasitology Today (Personal ed.)*, 15(12):497–501, December 1999.

[42] S Gupta, K Trenholme, R M Anderson, and K P Day. Antigenic diversity and the transmission dynamics of Plasmodium falciparum. *Science*, 263(5149):961–963, February 1994.

[43] S Gupta, M C Maiden, I M Feavers, S Nee, R M May, and R M Anderson. The maintenance of strain structure in populations of recombining infectious agents. *Nature Medicine*, 2(4):437–442, April 1996.

[44] S S Gupta, J J Swinton, and R M RM Anderson. Theoretical studies of the effects of heterogeneity in the parasite population on the transmission dynamics of malaria. *Proceedings of the Royal Society B: Biological Sciences*, 256(1347):231–238, June 1994.

[45] Sunetra Gupta. Parasite immune escape: new views into host-parasite interactions. *Current Opinion in Microbiology*, 8(4):6–6, August 2005.

[46] Jiang H, Li N, Gopalan, V, Zilversmit MM, Varma S, Nagarajan, V, Li J, Mu J, Hayton K, Henschen B, Yi M, Stephens R, McVean G, Awadalla P, Wellems TE, and Su XZ. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. *Genome Biology*, 12(4):R33–R33, April 2011.

[47] Daniel L Hartl and Andrew G Clark. *Principles of Population Genetics*. Sinauer Associates, Sunderland, 3 edition, 2007.

[48] B Haubold and R R Hudson. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. *Bioinformatics (Oxford, England)*, 16(9):847–848, September 2000.

[49] Richard R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, February 2002.

[50] Ronan Jambou, Eric Legrand, Makhtar Niang, Nimol Khim, Pharath Lim, Béatrice Volney, Marie Thérèse Ekala, Christiane Bouchier, Philippe Esterre, Thierry Fandeur, and Odile Mercereau-Puijalon. Resistance of Plasmodium falciparum field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. *Lancet*, 366 (9501):1960–1963, December 2005.

[51] Daniel C DC Jeffares, Arnab A Pain, Andrew A Berry, Anthony V AV Cox, James J Stalker, Catherine E CE Ingle, Alan A Thomas, Michael A MA Quail, Kyle K Siebenthall, Anne-Catrin AC Uhlemann, Sue S Kyes, Sanjeev S Krishna, Chris C Newbold, Emmanouil T ET Dermitzakis, and Matthew M Berriman. Genome variation and evolution of the malaria parasite Plasmodium falciparum. *Nature Genetics*, 39(1): 120–125, January 2007.

[52] Deirdre A Joy, Xiaorong Feng, Jianbing Mu, Tetsuya Furuya, Kesinee Chotivanich, Antoniana U Krettli, May Ho, Alex Wang, Nicholas J White, Edward Suh, Peter Beerli, and Xin-Zhuan Su. Early origin and recent expansion of Plasmodium falciparum. *Science*, 300(5617):318–321, April 2003.

[53] M Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, February 1968.

[54] M Kimura and J F CROW. THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION. *Genetics*, 49:725–738, April 1964.

[55] M M Kimura and T T Ohta. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics*, 61(3):763–771, March 1969.

[56] J L King and T H Jukes. Non-Darwinian evolution. *Science*, 164(3881):788–798, May 1969.

[57] M Korsinczky, N Chen, B Kotecka, A Saul, K Rieckmann, and Q Cheng. Mutations in Plasmodium falciparum cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site. *Antimicrobial Agents and Chemotherapy*, 44(8):2100–2108, August 2000.

[58] Sabrina Krief, Ananias A Escalante, M Andreina Pacheco, Lawrence Mugisha, Claudine André, Michel Halbwax, Anne Fischer, Jean-Michel Krief, John M Kasenene, Mike Crandfield, Omar E Cornejo, Jean-Marc Chavatte, Clara Lin, Franck Letourneur, Anne Charlotte Grüner, Thomas F McCutchan, Laurent Rénia, and Georges Snounou.

On the diversity of malaria parasites in African apes and the origin of Plasmodium falciparum from Bonobos. *PLoS Pathogens*, 6(2):e1000765–e1000765, January 2010.

[59] S Kyes, P Horrocks, and C Newbold. Antigenic variation at the infected red cell surface in malaria. *Annual Review of Microbiology*, 55:673–707, 2001.

[60] Gregory I GI Lang and Andrew W AW Murray. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics*, 178(1):67–82, January 2008.

[61] Thomas Lavstsen, Pamela Magistrado, Cornelus C Hermsen, Ali Salanti, Anja T R Jensen, Robert Sauerwein, Lars Hviid, Thor G Theander, and Trine Staalsoe. Expression of Plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans. *Malaria Journal*, 4(1):21–21, 2005.

[62] Pharath Lim, Alisa P Alker, Nimol Khim, Naman K Shah, Sandra Incardona, Socheat Doung, Poravuth Yi, Denis Mey Bouth, Christiane Bouchier, Odile Mercereau Puijalon, Steven R Meshnick, Chansuda Wongsrichanalai, Thierry Fandeur, Jacques Le Bras, Pascal Ringwald, and Frédéric Ariey. Pfmdr1 copy number and arteminisin derivatives combination therapy failure in falciparum malaria in Cambodia. *Malaria Journal*, 8:11, 2009.

[63] Weimin Liu, Yingying Li, Gerald H Learn, Rebecca S Rudicell, Joel D Robertson, Brandon F Keele, Jean-Bosco N Ndjango, Crickette M Sanz, David B Morgan, Sabrina Locatelli, Mary K Gonder, Philip J Kranzusch, Peter D Walsh, Eric Delaporte, Eitel Mpoudi-Ngole, Alexander V Georgiev, Martin N Muller, George M Shaw, Martine Peeters, Paul M Sharp, Julian C Rayner, and Beatrice H Hahn. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*, 467(7314):420–425, September 2010.

[64] Michael Lynch. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, 182(1):295–301, May 2009.

[65] Michael M Lynch. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Molecular Biology and Evolution*, 25(11):2409–2419, November 2008.

[66] RB Malaria. *Focus on Senegal*. Progress & Impact Series, 2010.

[67] Roll Back Malaria. Roll Back Malaria (RBM) Partnership - Malaria Key Facts. URL http://www.rollbackmalaria.org/keyfacts.html.

[68] Africa Region World Bank Malaria Implementation Resource Team. The World Bank Booster Program for Malaria Control in Africa, December 2007. URL http://siteresources.worldbank.org/EXTAFRBOOPRO/Resources/MALARIAREPORTfinalLOWRES.pdf.

[69] malERA Consultative Group on Monitoring, Evaluation, and Surveillance, Pedro L Alonso, Hoda Youseff Atta, Chris Drakeley, Thomas Eisele, Simon I Hay, Mario Henry Rodríguez Lúpez, Sylvia Meek, Richard Steketee, and Laurence Slutsker. A research agenda for malaria eradication: monitoring, evaluation, and surveillance. *PLoS Medicine*, 8(1):e1000400, 2011.

[70] Magnus Manske, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, Abdoulaye Djimde, Ogobara Doumbo, Issaka Zongo, Jean Bosco Ouédraogo, Pascal Michon, Ivo Mueller, Peter Siba, Alexis Nzila, Steffen Borrmann, Steven M Kiara, Kevin Marsh, Hongying Jiang, Xin-Zhuan Su, Chanaki Amaratunga, Rick Fairhurst, Duong Socheat, François Nosten, Mallika Imwong, Nicholas J White, Mandy Sanders, Elisa Anastasi, Dan Alcock, Eleanor Drury, Samuel Oyola, Michael A Quail, Daniel J Turner, Valentin Ruano-Rubio, Dushyanth Jyothi, Lucas Amenga-Etego, Christina Hubbart, Anna Jeffreys, Kate Rowlands, Colin Sutherland, Cally Roper, Valentina Mangano, David Modiano, John C Tan, Michael T Ferdig, Alfred Amambua-Ngwa, David J Conway, Shannon Takala-Harrison, Christopher V Plowe, Julian C Rayner, Kirk A Rockett, Taane G Clark, Chris I Newbold, Matthew Berriman, Bronwyn MacInnis, and Dominic P Kwiatkowski. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379, July 2012.

[71] J H McDonald and M Kreitman. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654, June 1991.

[72] D A Milner, Jr, S B Kamiza, C P Dzamalala, and V Vanguri. Continuing study of pediatric fatal malaria in Blantyre, Malawi. *International Journal for Parasitology*, 11: 35, February 2008.

[73] Danny A DA Milner, Charles P CP Dzamalala, N George NG Liomba, Malcolm E ME Molyneux, and Terrie E TE Taylor. Sampling of supraorbital brain tissue after death: improving on the clinical diagnosis of cerebral malaria. *The Journal of Infectious Diseases*, 191(5):805–808, March 2005.

[74] Victor A VA Mobegi, Kovana M KM Loua, Ambroise D AD Ahouidi, Judith J Satoguina, Davis C DC Nwakanma, Alfred A Amambua-Ngwa, and David J DJ Conway. Population genetic structure of Plasmodium falciparum across a region of diverse endemicity in West Africa. *Malaria Journal*, 11:223–223, January 2012.

[75] Jacqui Montgomery, Danny A Milner, Man Tsuey Tse, Alfred Njobvu, Kondwani Kayira, Charles P Dzamalala, Terrie E Taylor, Stephen J Rogerson, Alister G Craig, and Malcolm E Molyneux. Genetic analysis of circulating and sequestered populations of Plasmodium falciparum in fatal pediatric malaria. *The Journal of Infectious Diseases*, 194(1):115–122, July 2006.

[76] Bruno Moonen, Justin M Cohen, Robert W Snow, Laurence Slutsker, Chris Drakeley, David L Smith, Rabindra R Abeyasinghe, Mario Henry Rodriguez, Rajendra Maharaj, Marcel Tanner, and Geoffrey Targett. Operational strategies to achieve and maintain malaria elimination. *Lancet*, 376(9752):1592–1603, November 2010.

[77] Jianbing Mu, Philip Awadalla, Junhui Duan, Kate M Mcgee, Deirdre A Joy, Gilean A T Mcvean, and Xin-Zhuan Su. Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biology*, 3(10):e335–e335, October 2005.

[78] Jianbing Mu, Rachel A Myers, Hongying Jiang, Shengfa Liu, Stacy Ricklefs, Michael Waisberg, Kesinee Chotivanich, Polrat Wilairatana, Srivicha Krudsood, Nicholas J White, Rachanee Udomsangpetch, Liwang Cui, May Ho, Fengzhen Ou, Haibo Li, Jianping Song, Guoqiao Li, Xinhua Wang, Suon Seila, Sreng Sokunthea, Duong Socheat, Daniel E Sturdevant, Stephen F Porcella, Rick M Fairhurst, Thomas E Wellems, Philip Awadalla, and Xin-Zhuan Su. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genetics*, 42(3):268–271, March 2010.

[79] Jianbing J Mu, Philip P Awadalla, Junhui J Duan, Kate M KM McGee, Jon J Keebler, Karl K Seydel, Gilean A T GA McVean, and Xin-Zhuan XZ Su. Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nature Genetics*, 39(1):126–130, January 2007.

[80] Felista Mwingira, Gamba Nkwengulila, Sonja Schoepflin, Deborah Sumari, Hans-Peter Beck, Georges Snounou, Ingrid Felger, Piero Olliaro, and Kefas Mugittu. Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa. *Malaria Journal*, 10:79, 2011.

[81] Themba Mzilahowa, Philip J Mccall, and Ian M Hastings. "Sexual" population structure and genetics of the malaria agent P. falciparum. *PLoS ONE*, 2(7):e613, 2007.

[82] Shalini Nair, Alan Brockman, Lucy Paiphun, François Nosten, and Tim J C Anderson. Rapid genotyping of loci involved in antifolate drug resistance in Plasmodium falciparum by primer extension. *International Journal for Parasitology*, 32(7):852–858, June 2002.

[83] José A Nájera, Matiana González-Silva, and Pedro L Alonso. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS Medicine*, 8(1):e1000412, 2011.

[84] Mamadou O Ndiath, Catherine Mazenot, Ablaye Gaye, Lassana Konate, Charles Bouganali, Ousmane Faye, Cheikh Sokhna, and Jean-François Trape. Methods to collect Anopheles mosquitoes and evaluate malaria transmission: a comparative study in two villages in Senegal. *Malaria Journal*, 10:270, September 2011.

[85] Daouda Ndiaye, Vishal Patel, Allison Demas, Michele LeRoux, Omar Ndir, Souleymane Mboup, Jon Clardy, Viswanathan Lakshmanan, Johanna P Daily, and Dyann F Wirth. A non-radioactive DAPI-based high-throughput in vitro assay to assess Plasmodium falciparum responsiveness to antimalarials–increased sensitivity of P. falciparum to chloroquine in Senegal. *The American Journal of Tropical Medicine and Hygiene*, 82(2):228–230, February 2010.

[86] Daniel E Neafsey, Stephen F Schaffner, Sarah K Volkman, Daniel Park, Philip Montgomery, Danny A Milner, Amanda Lukens, David Rosen, Rachel Daniels, Nathan Houde, Joseph F Cortese, Erin Tyndall, Casey Gates, Nicole Stange-Thomann, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Marcelo U Ferreira, Sandra do Lago Moraes, Aditya P Dash, Chetan E Chitnis, Roger C Wiegand, Daniel L Hartl, Bruce W Birren, Eric S Lander, Pardis C Sabeti, and Dyann F Wirth. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. *Genome Biology*, 9(12):R171, 2008.

[87] Standwell C Nkhoma, Shalini Nair, Salma Al-Saai, Elizabeth Ashley, Rose McGready, Aung P Phyo, François Nosten, and Tim J C Anderson. Population genetic correlates of declining transmission in a human pathogen. *Molecular Ecology*, 22(2):273–85, November 2012.

[88] Michael F Ofori, Daniel Dodoo, Trine Staalsoe, Jørgen A L Kurtzhals, Kwadwo Koram, Thor G Theander, Bartholomew D Akanmori, and Lars Hviid. Malaria-induced acquisition of antibodies to Plasmodium falciparum variant surface antigens. *Infection and Immunity*, 70(6):2982–2988, June 2002.

[89] T Ohta. JSTOR: Annual Review of Ecology and Systematics, Vol. 23 (1992), pp. 263-286. *Annual Review of Ecology and Systematics*, 1992.

[90] Daniel J Park, Amanda K Lukens, Daniel E Neafsey, Stephen F Schaffner, Hsiao-Han Chang, Clarissa Valim, Ulf Ribacke, Daria Van Tyne, Kevin Galinsky, Meghan Galligan, Justin S Becker, Daouda Ndiaye, Souleymane Mboup, Roger C Wiegand, Daniel L Hartl, Pardis C Sabeti, Dyann F Wirth, and Sarah K Volkman. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *Proceedings of the National Academy of Sciences of the United States of America*, 109(32):13052–13057, August 2012.

[91] Daniel J Park, Amanda K Lukens, Daniel E Neafsey, Stephen F Schaffner, Hsiao-Han Chang, Clarissa Valim, Ulf Ribacke, Daria Van Tyne, Kevin Galinsky, Meghan Galligan, Justin S Becker, Daouda Ndiaye, Souleymane Mboup, Roger C Wiegand, Daniel L Hartl, Pardis C Sabeti, Dyann F Wirth, and Sarah K Volkman. Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. *Proceedings of the National Academy of Sciences of the United States of America*, 109(32):13052–13057, August 2012.

[92] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190–e190, December 2006.

[93] Jennifer J Peters, Elizabeth E Fowler, Michelle M Gatton, Nanhua N Chen, Allan A Saul, and Qin Q Cheng. High diversity and rapid changeover of expressed var genes during the acute phase of Plasmodium falciparum infections in human volunteers. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (16):10689–10694, August 2002.

[94] D S Peterson, D Walliker, and T E Wellems. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23):9114–9118, December 1988.

[95] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

[96] F Prugnolle, B Ollomo, P Durand, E Yalcindag, C Arnathau, E Elguero, A Berry, X Pourrut, J-P Gonzalez, D Nkoghe, J Akiana, D Verrier, E Leroy, F J Ayala, and F Renaud. African monkeys are infected by Plasmodium falciparum nonhuman primate-specific strains. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):11948–11953, July 2011.

[97] M B Reed, K J Saliba, S R Caruana, K Kirk, and A F Cowman. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium falciparum. *Nature*, 403(6772):906–909, February 2000.

[98] Ulf U Ribacke, Bobo W BW Mok, Valtteri V Wirta, Johan J Normark, Joakim J Lundeberg, Fred F Kironde, Thomas G TG Egwang, Peter P Nilsson, and Mats M Wahlgren. Genome wide gene amplifications and deletions in Plasmodium falciparum. *Molecular and Biochemical Parasitology*, 155(1):12–12, September 2007.

[99] Stephen M Rich, Fabian H Leendertz, Guang Xu, Matthew LeBreton, Cyrille F Djoko, Makoah N Aminake, Eric E Takang, Joseph L D Diffo, Brian L Pike, Benjamin M Rosenthal, Pierre Formenty, Christophe Boesch, Francisco J Ayala, and Nathan D Wolfe. The origin of malignant malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 2009.

[100] C Roper, I M Elhassan, L Hviid, H Giha, W Richardson, H Babiker, G M Satti, T G Theander, and D E Arnot. Detection of very low level Plasmodium falciparum infections using the nested polymerase chain reaction and a reassessment of the epidemiology of unstable malaria in Sudan. *The American Journal of Tropical Medicine and Hygiene*, 54(4):325–331, April 1996.

[101] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet,

Stephen F Schaffner, Eric S Lander, International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, Todd A Johnson, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Bir-

ren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, and John... Wallenburg. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449 (7164):913–918, October 2007.

[102] P Sjödin. On the Meaning and Existence of an Effective Population Size. *Genetics*, 169(2):1061–1070, February 2005.

[103] J M Smith, N H Smith, and M O'Rourke. How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, 90(10):4384 – 4388, May 1993.

[104] John A Stamatoyannopoulos, Ivan Adzhubei, Robert E Thurman, Gregory V Kryukov, Sergei M Mirkin, and Shamil R Sunyaev. Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393–395, April 2009.

[105] Kathryn N Suh, Kevin C Kain, and Jay S Keystone. Malaria. *Canadian Medical Association Journal*, 170(11):1693–1702, May 2004.

[106] F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, November 1989.

[107] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, May 2005.

[108] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *The American Journal of Human Genetics*, 79(1):12–12, July 2006.

[109] T E Taylor, WJJ Fu, R A Carr, R O Whitten, J G Mueller, N G Fosiko, S Lewallen, N G Liomba, and M E Molyneux. Differentiating the pathologies of cerebral malaria by postmortem parasite counts. *Nature Medicine*, 10(2):143–145, February 2004.

[110] Anne-Catrin Uhlemann, Angus Cameron, Ursula Eckstein-Ludwig, Jorge Fischbarg, Pavel Iserovich, Felipe A Zuniga, Malcolm East, Anthony Lee, Leo Brady, Richard K Haynes, and Sanjeev Krishna. A single amino acid residue can determine the sensitivity of SERCAs to artemisinins. *Nature structural & molecular biology*, 12(7):628–629, July 2005.

[111] Daria Van Tyne, Daniel J Park, Stephen F Schaffner, Daniel E Neafsey, Elaine Angelino, Joseph F Cortese, Kayla G Barnes, David M Rosen, Amanda K Lukens, Rachel F Daniels, Danny A Milner, Charles A Johnson, Ilya Shlyakhter, Sharon R

Grossman, Justin S Becker, Daniel Yamins, Elinor K Karlsson, Daouda Ndiaye, Ousmane Sarr, Souleymane Mboup, Christian Happi, Nicholas A Furlotte, Eleazar Eskin, Hyun Min Kang, Daniel L Hartl, Bruce W Birren, Roger C Wiegand, Eric S Lander, Dyann F Wirth, Sarah K Volkman, and Pardis C Sabeti. Identification and Functional Validation of the Novel Antimalarial Resistance Locus PF10_0355 in Plasmodium falciparum. *PLoS Genetics*, 7(4):e1001383, April 2011.

[112] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4(3):e72, March 2006.

[113] Sarah K Volkman, Pardis C Sabeti, David DeCaprio, Daniel E Neafsey, Stephen F Schaffner, Danny A Milner, Johanna P Daily, Ousmane Sarr, Daouda Ndiaye, Omar Ndir, Soulyemane Mboup, Manoj T Duraisingh, Amanda Lukens, Alan Derr, Nicole Stange-Thomann, Skye Waggoner, Robert Onofrio, Liuda Ziaugra, Evan Mauceli, Sante Gnerre, David B Jaffe, Joanne Zainoun, Roger C Wiegand, Bruce W Birren, Daniel L Hartl, James E Galagan, Eric S Lander, and Dyann F Wirth. A genome-wide map of diversity in Plasmodium falciparum. *Nature Genetics*, 39(1):113–119, January 2007.

[114] R S Waples. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391, February 1989.

[115] B S Weir and C C Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370, November 1984.

[116] V A VA White, S S Lewallen, N N Beare, K K Kayira, R A RA Carr, and T E TE Taylor. Correlation of retinal haemorrhages with brain haemorrhages in children dying of cerebral malaria in Malawi. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 95(6):618–621, November 2001.

[117] Valerie A White, Susan Lewallen, Nicholas A V Beare, Malcolm E Molyneux, Terrie E Taylor, and Vasee Moorthy. Retinal Pathology of Pediatric Cerebral Malaria in Malawi. *PLoS ONE*, 4(1):e4317–e4317, January 2009.

[118] C M Wilson, A E Serrano, A Wasley, M P Bogenschutz, A H Shankar, and D F Wirth. Amplification of a gene related to mammalian mdr genes in drug-resistant Plasmodium falciparum. *Science*, 244(4909):1184–1186, June 1989.

[119] C M Wilson, S K Volkman, S Thaithong, R K Martin, D E Kyle, W K Milhous, and D F Wirth. Amplification of pfmdr 1 associated with mefloquine and halofantrine resistance in Plasmodium falciparum from Thailand. *Molecular and Biochemical Parasitology*, 57 (1):151–160, January 1993.

[120] Paul E Wilson, Alisa P Alker, and Steven R Meshnick. Real-time PCR methods for monitoring antimalarial drug resistance. *Trends in Parasitology*, 21(6):278–283, June 2005.

[121] K H Wolfe, P M Sharp, and W H Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285, January 1989.

[122] World Health Organization. Severe falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1(94):S1–90, April 2000.

[123] World Health Organization. *World Malaria Report 2010 (Who Global Malaria Programme)*. World Health Organization, May 2011.

[124] World Health Organization. World malaria report 2011. *Geneva: World Health Organization*, 246, 2011.

[125] World Health Organization. *World Malaria Report 2012*. Geneva: World Health Organization, December 2012.

[126] S Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, March 1931.

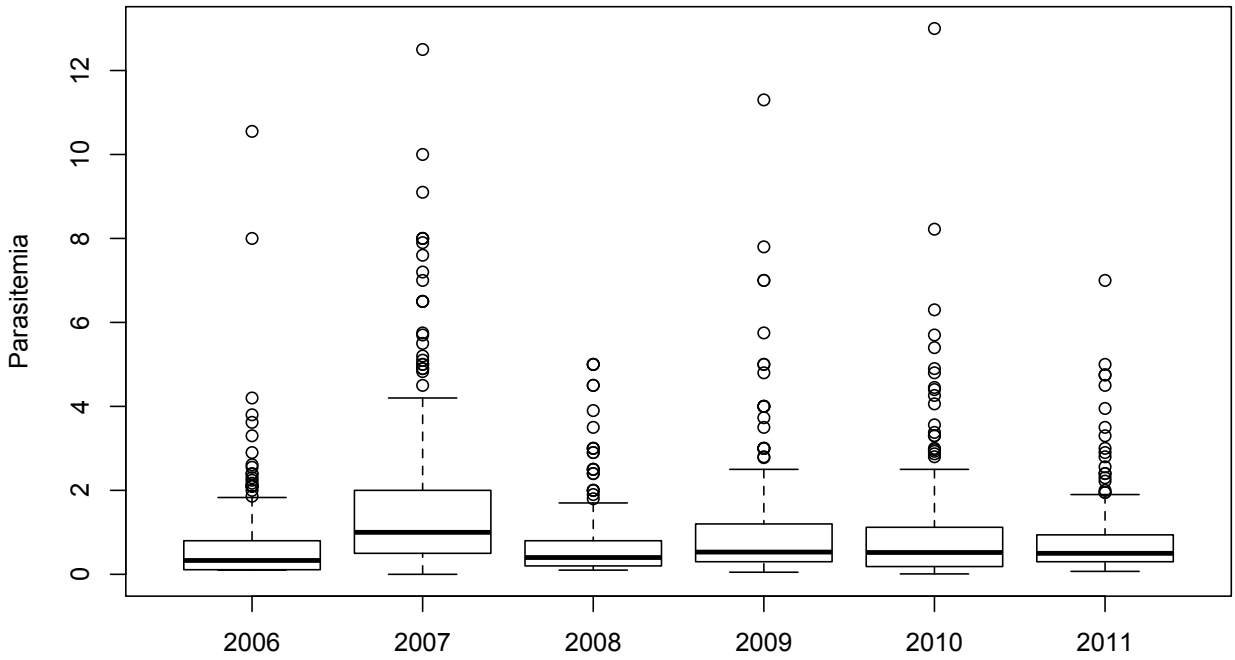
[127] Jing Yuan, Ken Chih-Chien Cheng, Ronald L Johnson, Ruili Huang, Sittiporn Pattaradilokrat, Anna Liu, Rajarshi Guha, David A Fidock, James Inglese, Thomas E Wellems, Christopher P Austin, and Xin-Zhuan Su. Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. *Science*, 333(6043): 724–729, August 2011.

[128] Zhiyong Zhou, Amanda C Poe, Josef Limor, Katharine K Grady, Ira Goldman, Andrea M McCollum, Ananias A Escalante, John W Barnwell, and Venkatachalam Udhayakumar. Pyrosequencing, a high-throughput method for detecting single nucleotide polymorphisms in the dihydrofolate reductase and dihydropteroate synthetase genes of Plasmodium falciparum. *Journal of Clinical Microbiology*, 44(11):3900–3910, November 2006.

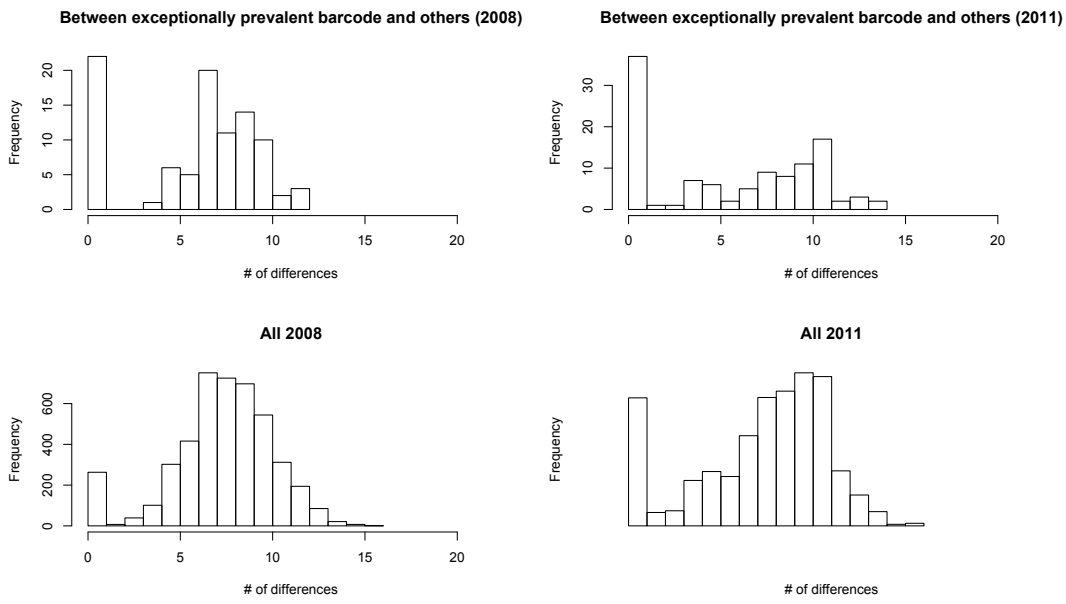
Appendices

	SenT028.09	SenT142.09	SenT029.09	SenT132.09	SenT061.09	SenT072.09	Dd2#1	Dd2#2
SenT028.09	0.0	1.7	19.4	19.5	18.7	18.7	28.2	27.5
SenT142.09	1.7	0.0	19.3	19.5	19.2	18.8	28.7	27.7
SenT029.09	19.4	19.3	0.0	1.5	19.0	19.5	28.4	29.1
SenT132.09	19.5	19.5	1.5	0.0	19.3	19.1	28.6	28.7
SenT061.09	18.7	19.2	19.0	19.3	0.0	1.9	27.4	28.3
SenT072.09	18.7	18.8	19.5	19.1	1.9	0.0	28.5	28.8
Dd2#1	28.2	28.7	28.4	28.6	27.4	28.5	0.0	2.0
Dd2#2	27.5	27.7	29.1	28.7	28.3	28.8	2.0	0.0

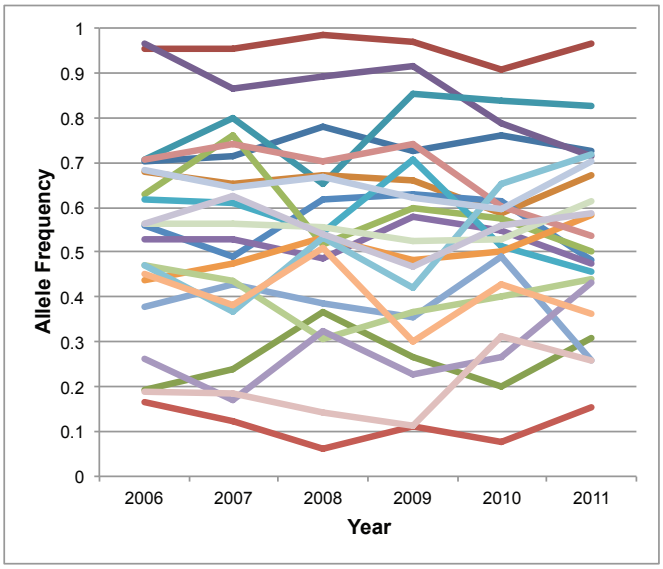
Supplementary Figure 1: The percent differences between hybridized biological replicates and samples with identical barcodes. Array analysis shows that the percentage of SNP differences between samples with identical barcodes is similar to those seen in biological replicates, suggesting that samples with identical barcodes are nearly genetically identical.



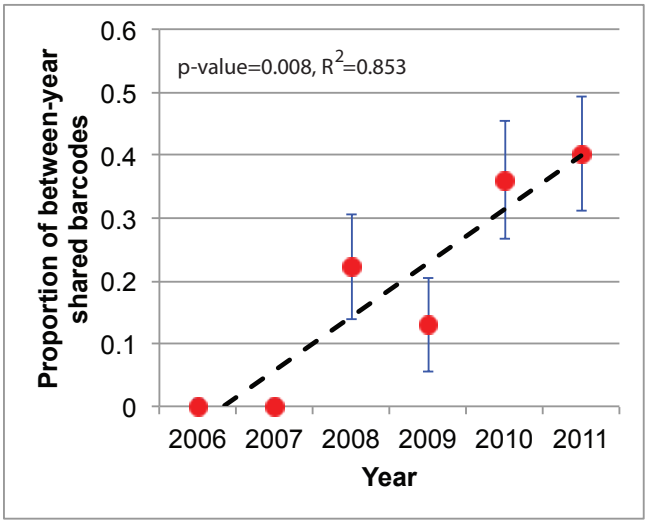
Supplementary Figure 2: Parasitemia variation by year. Kruskal-Wallis rank-sum test indicates variance between years (p-value $<$ 2.2e-16); however, there is no decreasing trend over time.



Supplementary Figure 3: Pairwise differences between two exceptionally prevalent barcodes and other barcodes. The number of pairwise differences between the exceptionally prevalent barcode and other barcodes is not significantly higher than the number of pairwise differences among all the barcodes in the same year.



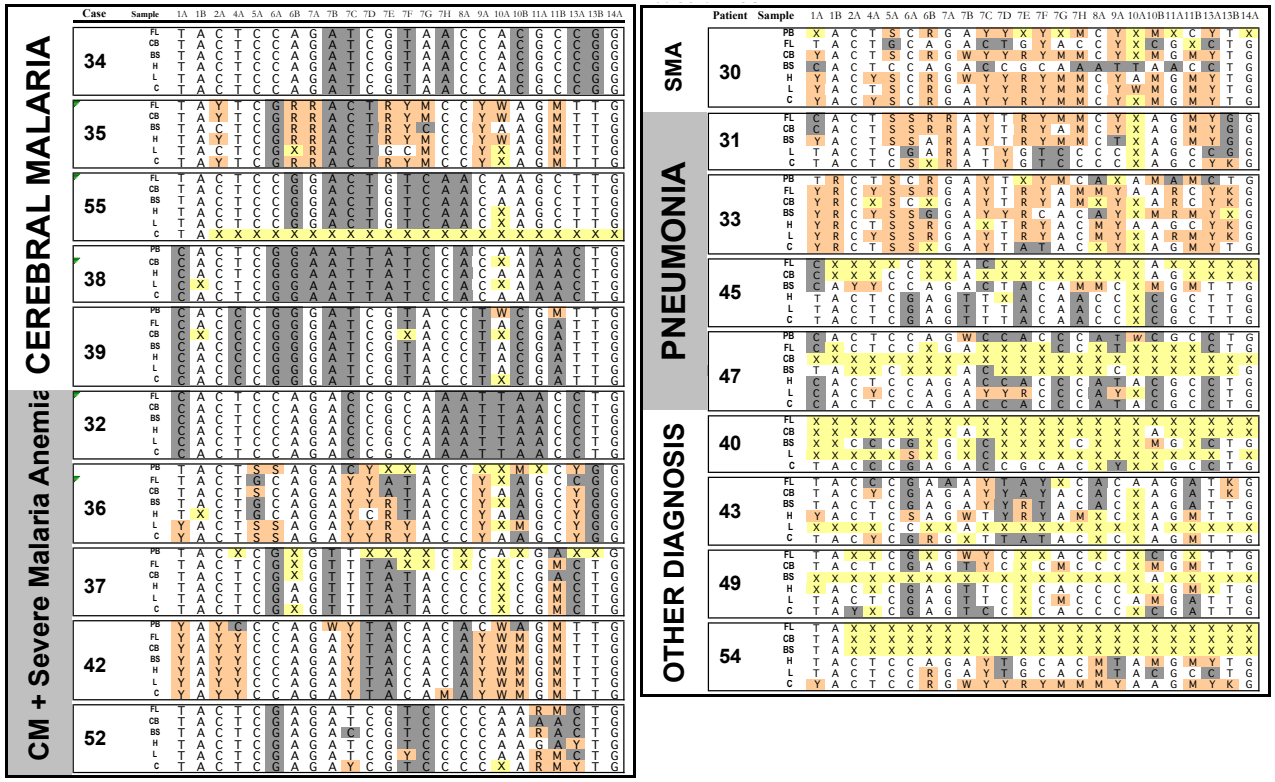
Supplementary Figure 4: Changes in allele frequencies between transmission seasons. Each colored line shows the year-to-year variability in allele frequency for each non-fixed SNP. Each colored line shows the year-to-year variability in allele frequency for each non-fixed SNP. Allele frequencies fluctuate substantially across years, suggesting high random genetic drift and low effective population size.



Supplementary Figure 5: Proportion of between-year shared barcodes. Proportion of between-year shared barcodes increased significantly. The error bars show 95% confidence interval of mean ($\pm$1.96 SE).

		CM	CM/SMA	Pneumonia	Other Cause	P-Value
Autopsy	*msp-1 & -2*	1.9	2.9	3.3	2.8	<0.001
	Barcode	1.6	3.6	6.5	3.4	<0.001
		Retinopathy Positive		Retinopathy Negative		
Clinical	Barcode	3.5		5.8		0.018

Supplementary Figure 6: Comparison of the number of *msp*-1 and *msp*-2 alleles found in peripheral blood and tissues of patients with different diagnosis with the number of heterozygous calls found using the molecular barcode demonstrates low complexity in cerebral malaria. Statistical comparison was made by ANOVA. Comparison of the average number of heterozygous calls found using the molecular barcode in the clinical peripheral blood samples shows a significant difference between retinopathy positive and retinopathy negative. The slightly higher average value of the retinopathy positives is due to the inclusion of patients who may be CM and SMA. Statistical comparison was made by *t*-test of means.



Supplementary Figure 7: The complete molecular barcodes for 19 patients with pathological evidence of cerebral malaria at autopsy. Sample types include peripheral blood (PB), frontal lobe (FB), cerebellum (CB), brainstem (BS), heart (H), lung (L), and colon (C). Major alleles are white background and the minor alleles gray background. The frequency of major and minor alleles for the population of Malawi parasites examined in this study were significantly different from those seen in the global population (Table 5.3.2). For a heterozygous call, the IUPAC code is used and BOTH bases denoted by the code are present (orange/light-gray background). Failed reactions are shown as "X".

Supplementary Table 1: The number of mixed and single infections in each year.

	Single	Mixed	Total
2006	90	41	131
2007	54	26	80
2008	95	13	108
2009	77	15	92
2010	100	25	125
2011	112	26	138

Supplementary Table 2: Variance effective population size.

	Moment		Likelihood	
	Mean	95% CI	Mean	95% CI
All samples				
2006-2007	ND*	(70, ND)	ND	(226, ND)
2007-2008	18	(7, 46)	19	(9, 49)
2008-2009	24	(9, 67)	29	(12, 90)
2009-2010	16	(7, 36)	18	(9, 42)
2010-2011	9	(4, 16)	10	(6, 18)
Ignoring duplicate barcodes				
2006-2007	ND	(116, ND)	ND	(132, ND)
2007-2008	82	(16, ND)	85	(19, ND)
2008-2009	138	(21, ND)	197	(27, ND)
2009-2010	59	(14, ND)	77	(18, ND)
2010-2011	240	(22, ND)	214	(24, ND)

* ND represents “Not Determinable”.

Supplementary Table 3: Human genotyping data. TaqMan probes from Broad Institute set

Sample Name	rs1009806	rs898500	rs1000797	rs1000026	rs1000005	rs10242744	rs1036689	rs1000203	rs1037439	rs242076	rs1012315	rs1015939	rs1000158	rs1571256	rs1025412	rs1000192	rs10775365	rs1000053	rs1000121	rs10513695
SenT005.11	TT	AG	GG	CC	CG	AG	GG	TT	AG	AG	AG	CC	GG	AA	GG	GG	GG	CC	CT	GT
SenT014.11	CT	AG	-	CC	CG	AA	GG	TT	AG	AG	AA	-	GG	AA	GG	GG	AG	CC	CT	GT
SenT018.11	TT	AA	-	CC	GG	GG	GG	TT	AA	AA	AG	-	AG	AA	GG	GG	GG	CC	CC	GG
SenT019.11	TT	AA	GG	CC	GG	AG	AG	TT	AA	AA	AA	CC	GG	AA	AA	AG	GG	CT	CT	GT
SenT021.11	TT	AA	GG	CT	GG	AA	AG	TT	GG	AA	AG	-	AG	AA	AG	GG	-	CC	CT	TT
SenT026.11	TT	AG	GG	CC	CG	GG	GG	TT	GG	AG	GG	-	GG	AA	AG	GG	GG	CT	TT	GT
SenT054.11	TT	AG	GG	CT	CG	AG	AG	TT	AA	AG	AA	-	GG	AA	GG	GG	GG	CC	CT	GT
SenT064.11	TT	AG	GG	CC	CG	GG	GG	TT	AA	-	AG	CC	-	AA	AG	GG	GG	CC	CT	TT
SenT078.11	TT	AA	GG	CT	GG	AG	AG	TT	AG	AG	GG	CC	GG	AA	AG	AG	GG	CC	TT	TT
SenT095.11	TT	AG	-	CC	GG	AG	GG	TT	AA	AA	AG	CC	GG	AA	AG	GG	GG	CC	CT	TT
SenT098.11	TT	GG	GG	CT	GG	AA	GG	TT	AG	AG	GG	-	AG	AA	AG	GG	GG	CT	TT	GT
SenT101.11	CT	AG	-	CT	CG	AG	GG	TT	AA	AG	AA	CC	AG	AA	AA	GG	GG	CC	TT	TT
SenT103.11	TT	AG	-	CC	GG	GG	AG	TT	AA	AG	AG	CC	GG	AA	GG	GG	GG	CC	TT	TT
SenT104.11	TT	AG	-	CT	GG	AG	GG	TT	AG	AG	AG	-	GG	AA	GG	GG	GG	CC	CT	GT
SenT105.11	TT	AG	-	CC	GG	AG	AG	TT	GG	AA	GG	CC	AG	AG	GG	GG	GG	CC	TT	TT
SenT106.11	CT	AG	GG	CT	CC	AA	GG	TT	AG	AG	GG	CC	GG	AA	AG	AG	GG	CC	CT	TT
SenT110.11	TT	AG	GG	CT	CG	AA	GG	TT	AA	AG	AG	CC	AG	AA	GG	GG	GG	CC	TT	TT
SenT111.11	CT	AG	GG	CC	CG	GG	AG	TT	AA	AG	AG	TT	GG	AA	AA	AG	GG	CC	CC	TT
SenT117.11	TT	AG	GG	CT	GG	AA	AG	TT	GG	AG	GG	CC	AG	AA	AA	AG	AA	CC	-	TT
SenT122.11	TT	GG	GG	CT	GG	AG	GG	TT	AA	AG	AA	CC	AG	AA	AA	GG	GG	CC	CC	TT
SenT123.11	CT	GG	-	CT	GG	AA	GG	TT	AG	AA	AG	CC	AG	AA	AA	GG	GG	CT	TT	TT
SenT127.11	CT	AG	GG	CC	CG	AA	GG	CT	AA	AG	AG	-	AG	AA	AG	-	AG	CC	CC	TT
SenT132.11	TT	AG	GG	CT	GG	AG	-	TT	AA	AG	AG	-	GG	AA	AG	-	GG	CC	CC	GT
SenT134.11	CT	GG	GG	CC	GG	AG	GG	TT	GG	AG	AA	CT	GG	AA	GG	GG	GG	TT	CT	GT
SenT135.11	TT	AG	GG	CC	CG	AG	GG	TT	AA	AG	GG	CC	-	AA	AA	GG	GG	CC	CT	TT
SenT136.11	TT	AG	-	CT	CG	GG	AG	TT	AG	AA	AG	CC	GG	AA	AA	AG	AG	CT	TT	GG
SenT145.11	TT	AA	GG	CT	GG	AA	AG	TT	AA	AG	AA	CC	AG	AA	AG	AG	GG	CC	CT	TT
SenT146.11	CT	AA	-	CC	GG	GG	GG	TT	AA	AG	AG	CC	GG	AA	GG	GG	GG	CC	CT	TT

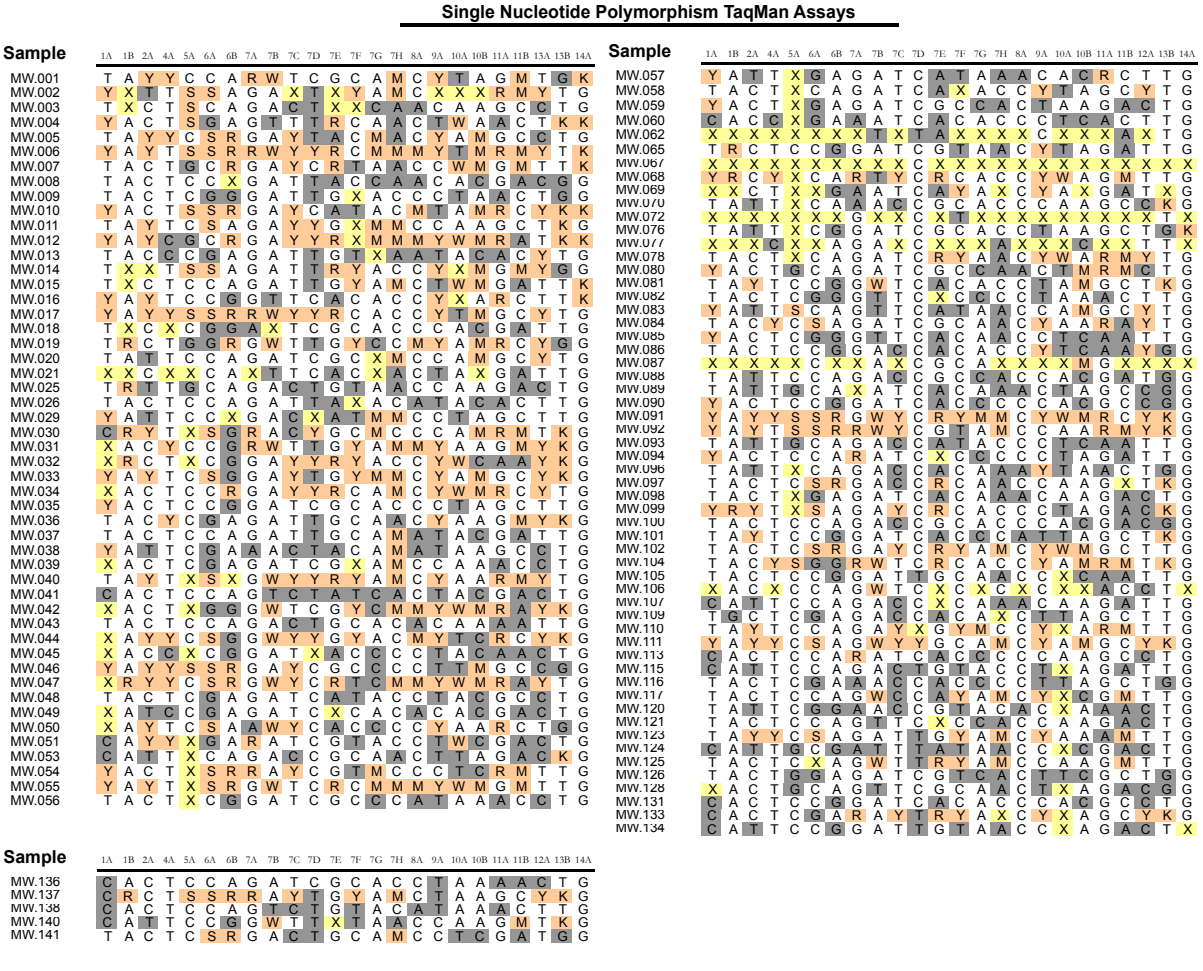
Supplementary Table 4: Human genotyping data. STR genotyping on an ABI 3130 Genetic Analyzer

STR Locus	Th 153.09	Th 093.09	Th 138.09	Th 142.09	Th 108.09	Th 109.09
D3S1358	16,18	15,16	17,18	15,-	15,16	15,16
vWA	15,18	19,20	13,16	15,20	15,16	15,16
FGA	24,26	21, (& 20?)	22,26	24,25(?)	22,23	21,24
Amelogenin	X,Y	X,Y	X,X	X,Y	X,Y	X,Y
D8S1179	11,14	12,-	15,16	14,15	15,-	13,16
D21S11	28,-	33,34	28,29	29,-	28,30	28,30
D18S51	11,15	12,19	16,17	17,20	17,18	17,18
D5S818	10,12	11,12	11,13	10,13	10,12	10,12
D13S317	12,13	12,14(?)	12,13	11,12	10,12	11,12
D7S820	10,11	11,-	10,-	8,10	9,13	9,-

Supplementary Table 5: csp sequences. The first group is unrelated parasites; the other three groups are clusters of identical barcodes.

3D7	A T T T A A A C A A A A T A C A A A A T T N C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th110_08	A T T T A A A G A A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th250_08	A T T T A A A G A C A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th045_08	A T T T G A A G N A A A T A C T T A A T T - C T C T T T C A A C T G A A C G G T C C C C A T G T A G T G T A A C T T
Th216_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th109_09	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th088_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th068_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th042_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th228_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th096_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th188_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th173_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th120_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th101_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th108_09	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th112_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th105_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th214_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th073_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th044_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th004_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th014_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th220_08	A T T T A A A G G A A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th096_06	A T T T A A A G A C A A T A A A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th134_06	A T T T A A A G A C A A T A A A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th127_08	A T T T A A A G A G A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T
Th246_08	A T T T A A A G A G A A T A C A A A A T T - C T C T T T C A A C T G A A T G G T C C C C A T G T A G T G T A A C T T

Supplementary Table 6: The barcodes for 112 consecutive patients with positive peripheral blood parasitaemia and sufficient DNA for performance of the molecular barcode are shown. The order of the patients displayed in the table is by study number and not by date.



Supplementary Table 7: Descriptions of the assays used in this study are shown including the derived major and minor allele frequencies for the Malawi data set (*) as well as codes for heterozygous allele calls (both alleles present) in the standard IUPAC format(**).

Assay Number	Assay ID	Major Allele*	Minor Allele	Heterozygous**
1	1A	T	C	Y
2	1B	A	G	R
3	2A	C	T	Y
4	4A	T	C	Y
5	5A	C	G	S
6	6A	C	G	S
7	6B	A	G	R
8	7A	G	A	R
9	7B	A	T	W
10	7C	T	C	Y
11	7D	C	T	Y
12	7E	G	A	R
13	7F	C	T	Y
14	7G	A	C	M
15	7H	C	A	M
16	8A	C	A	M
17	9A	C	T	Y
18	10A	A	T	W
19	10B	A	C	M
20	11A	G	A	R
21	11B	C	A	M
22	12A	T	C	Y
23	13B	T	G	K
24	14A	G	T	K